PQ-DAF：基于姿态驱动的质量控制数据增强用于数据稀缺的驾驶员分心检测

Haibin Sun [a] , Xinghui Song [a,*]

Abstract—**司机分心检测对于提高交通安全性和减少道路事故至关重要。然而，现有模型在实际场景中部署时常常面临泛化能力下降的问题。这一限制主要源于在实际环境中数据标注的高成本所导致的小样本学习挑战，以及训练数据集与目标部署环境之间的巨大域转移。为了解决这些问题，我们提出了一个姿态驱动的质量控制数据增强框架（PQ-DAF），该框架利用视觉-语言模型对样本进行筛选，经济高效地扩展训练数据并增强跨领域的鲁棒性。具体而言，我们采用一个渐进条件扩散模型（PCDMs）来准确捕捉关键的司机姿态特征，并综合生成多样化的训练示例。然后引入基于 CogVLM 视觉-语言模型的样本质量评估模块，根据置信度阈值筛选出低质量的合成样本，确保增强数据集的可靠性。大量实验表明，PQ-DAF 在小样本司机分心检测任务中显著提高了性能，在数据稀缺条件下实现了模型泛化能力的显著提升。**

Index Terms—driver distraction detection, data augmentation, diffusion models, pose-guided generation

I. 引言

D 河流 分心检测在增强交通安全和降低事故风险方面起着至关重要的作用。近年来，基于深度学习的方法已成为此任务的主流，并取得了显著的效果。然而，由于训练数据与真实世界部署场景之间的领域转移，这些模型在实际应用中常常遭遇性能显著下降。为了缓解这种数据转移，在部署时收集新的数据集是一种直观的解决方案。然而，数据收集和标注成本高且劳动密集，因此在少样本条件下提高模型鲁棒性显得尤为重要。

现有研究大致可以分为两个方向。第一个方向侧重于设计轻量且高效的分类模型，优化其在资源受限的边缘设备上的部署。例如，MTNet 将多维自适应特征提取模块与轻量特征融合集成，在 LDDB 基准上减少了 37 % 的参数，同时保持了 96.2 % 的准确率，尽管对于如手部动作之类的小目标有 8.3 % 的下降 [1] 。此外，RES-SE-CNN 结合残差网络与通道注意力，达到了 97.28 % 的准确率，同时显著降低了内存使用，使其适合车载应用 [2] 。然而，这些方法通常需要大规模标注数据集，并且在域转移时易受影响，限制了其泛化能力。

第二类研究通过数据增强或合成数据生成来缓解域移位，从而解决小样本问题 [3] 。在其他领域，图像变换、风格迁移和基于 GAN 的生成等技术已经被应用于扩充训练集并提高鲁棒性，但这些方法在驾驶员分心检测中仍未被充分探索 [4, 5] 。例如，Wang 等人 [6] 通过结合区域检测与图像增强在 StateFarm 数据集上达到了 96.97 % 的准确率，而 Hasan 等人 [7] 则采用 Sim2Real 策略进行跨域适应。尽管这些方法具有潜力，但它们通常面临不稳定的样本质量、高标注成本以及合成数据的可靠性低的问题，这限制了其在现实世界中的应用。

*Corresponding author. Email addresses: sdustsun@163.com (Haibin Sun), xinghuisong7026@163.com (Xinghui Song).
[a] College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, 266590, China.

为了解决这些挑战，我们提出了一种 基于姿态驱动的质量控制数据增强框架 (PQ-DAF) ，专门针对少样本驾驶员分心检测。首先，驾驶员姿态信息通过 DWpose 提取，并用作渐进条件扩散模型（PCDMs） [8] 的条件输入，以生成多样化的、高保真的训练样本，同时保持结构一致性。接着，视觉语言模型 CogVLM [9] 作为自动质量评估器，根据预定义的阈值过滤掉低置信度的合成样本，从而确保增强数据的可靠性。该方法消除了手动标注，使得数据集能够以较低成本扩展，并提升跨领域泛化能力。此外，我们研究了真实数据与合成数据的比例变化的效果。本文的主要贡献总结如下：

- 我们使用 DWpose 捕捉关键的人体姿势信息，并采用 PCDMs 合成多样化、高保真且具有结构一致性的驾驶员行为图像。
- 我们引入了一种基于 CogVLM 的自动质量过滤机制来去除低质量样本，从而提高扩增数据集的可靠性。
- 我们在多个真实的驾驶员行为数据集上进行了大量实验，证明我们的方法在小样本场景中始终提高了性能。

随着车辆的日益普及和道路交通的愈加复杂，交通安全问题变得越来越严重。根据世界卫生组织的统计数据，全球每年大约有 135 万人死于交通事故。研究表明，驾驶员分心是道路交通事故的重要原因之一。因此，检测驾驶员分心在道路安全领域具有重要意义。

在早期关于驾驶员注意力分散检测的研究中，研究人员主要依赖于传统的机器学习方法结合手工特征提取。例如，Billah 和 Rahman 等人 [10] 研究了使用手、唇和额头特征进行驾驶员分心检测，并采用了 K-近邻（KNN）算法，在其自建的数据集上达到了 81.50 % 至 81.67 % 的准确率。虽然这些方法在特定场景中取得了一定成果，但它们的特征提取过程复杂且固有局限性。随着深度学习的进步，以卷积神经网络（CNN）为基础的方法逐渐成为主流 [11–14] 。例如，轻量级 CNN 模型 OLCNNet [15] 在 StateFarm 数据集上实现了 89.53 % 的准确率。此外，由于 Transformer 模型的优异的长距离依赖建模能力，它们被引入到行为识别领域 [16, 17] 。BiRSwinT [18] 采用了基于 Swin Transformer 的双流结构，学习并融合全局和局部驾驶员动作特征，以实现细粒度的驾驶行为识别。然而，尽管这些方法在特定数据集上取得了良好的识别准确性，但由于训练数据与真实应用环境之间的域迁移问题，它们的性能在数据稀缺的场景中往往显著下降。这就需要收集新的数据集——这一过程由于数据获取和注释的困难而既复杂又昂贵。为了在数据稀缺的场景中增强模型性能，获取真实且多样的训练数据至关重要。然而，收集驾驶员行为数据常常面临高成本、严重的隐私问题和复杂场景等挑战，这使得在真实环境中获得全面的、高质量的异常行为样本变得困难重重。因此，探索更有效的数据扩展和增强方法，特别是能够模拟真实驾驶环境和行为的数据生成技术，已成为提升驾驶员分心检测模型在实际应用中性能

的关键。

A. 基于姿态的图像生成相关工作

姿态引导的图像生成在近年来取得了显著进展，其目标是在合成符合指定目标姿态的高质量图像。其核心思想是将目标姿态信息纳入生成过程中，使得合成结果既匹配源图像的风格又符合所需的姿态。早期研究将这一任务表述为利用条件生成对抗网络 (CGAN) [19] 的条件图像生成问题，其中源图像外观和目标姿态作为条件输入。然而，由于源姿态和目标姿态之间固有的不对齐，直接条件输入通常难以应对复杂的姿态变换。为了解决这一挑战，Def-GAN [20] 引入了一种可变形的 GAN，通过一组局部仿射变换来建模姿态变化，从而缓解不对齐问题。类似地，ADGAN [21] 使用纹理编码器提取身体部位的风格向量，并将其应用于 AdaIN 残差块来生成最终图像。这些方法通过设计专门用于复杂姿态变换的架构来专注于将源图像的风格转移到目标姿态，但在建模高度非线性变形方面仍面临限制。

近年来，扩散模型 [22–24] 作为强大的生成模型出现，与基于 GAN 的方法相比，产生了更优质且多样的图像。在无条件合成取得成功后，扩散模型迅速扩展到条件设置，包括姿势引导。早期的工作如 PIDM [25] 将 Stable Diffusion 改编以同时基于源图像和姿势图像进行条件生成，从而实现风格和姿势线索的有效融合。MGD [26] 进一步利用多模态条件——姿势图、服装草图和文本——在潜在空间中引导生成。渐进条件扩散模型 (PCDMS) [27] 通过逐步完善姿势和风格信息推进这一方向，而 HumanSD [28] 优化了复杂场景中的人类生成。最近的统一框架如 ImagPose [29] 和任务特定系统如 ImagDressing [30] 证明了基于扩散的条件生成可以灵活地处理姿势和外观定制，甚至在多阶段工作流程中 [31] 。然而，迄今为止，没有研究应用姿势引导的图像生成来检测驾驶员分心现象。鉴于分心行为与姿势模式密切相关，并且公共数据集很少与真实驾驶场景匹配，姿势引导生成提供了一种模拟各种驾驶员行为的有效方式。通过合成不同姿势配置，它可以丰富训练数据集，并有可能在实际部署中提高分心检测模型的鲁棒性和准确性。

II. 提出的方法

为了解决在驾驶员分心检测场景中由于手动标注成本高而导致的数据匮乏问题，我们提出了姿态驱动的质量控制数据增强框架（PQ-DAF），如图 ?? 所示。首先，我们从原始驾驶图像中提取人体关键点姿态作为结构先验。然后，我们使用渐进条件扩散模型（PCDMs）生成多样化、结构良好的伪样本，以扩展训练集。与传统数据增强或基于 GAN 的方法不同，PCDMs 提供了更强的建模能力和更高的稳定性。在给定的姿态约束下，它们生成语义清晰的图像，这些图像反映出一致的驾驶动作。

在样本生成之后，我们引入了一种基于视觉语言模型 CogVLM 的多模态推理的自动化质量控制机制。对于每一种姿势类别，我们定义了一个统一的语义提示，并评估生成的图像与该提示的匹配程度。我们为每个图文对计算一个置信度评分，并丢弃那些低于预设门槛的样本。此过滤步骤去除可能损害训练的噪声样本。最后，我们将保留的高质量伪样本与真实数据合并，并重新训练模型。通过将结构引导的增强与语义过滤相结合，PQ-DAF 有效地增强了训练数据，而无需额外的人工标注，从而在数据稀缺的情况下提高了模型性能。

我们提出了一种姿势条件图像生成模块，结合人体姿势提取和图像生成技术，以合成具有高语义一致性和视觉逼真度的训练样本，有效解决了因数据稀缺在驾驶员分心检测任务中导致的模型泛化能力不足的问题。具体来说，首先利用 DWpose 从源图像和目标图像中提取精细化的骨骼姿势图作为结构先验约束，然后将其输入到渐进条件扩散模型（PCDMs）的第二阶段生成器中，执行潜在空间条件扩散，生成语义合理和结构连贯的驾驶行为图像。

为了确保语义一致性并有效建模生成图像中的关键区域，模型输入由三个条件分支组成，如图 1 所示：（1）源图像及其对应的二值掩码；（2）配对的源图像和目标图像；(3) 源和目标姿势图。这些图像对沿宽度维度连接，以构建空间对齐的输入表示。为了防止由黑色像素引起的歧义，我们进一步引入一个单通道指示图（图中未显示），其中 0 和 1 分别表示被遮挡和未被遮挡的区域，帮助模型识别需要填充的区域。

对于特征编码，我们采用一个冻结的 DINOv2 [32] 编码器，从源图像中提取细粒度的语义表示，这些表示通过一个浅层的多层感知器（MLP）进一步投射到潜在空间。姿态图通过一个四层卷积姿态编码器进行处理，以保留空间结构并准确建模目标动作。多源信息被编码为三种类型的嵌入：融合图像嵌入 f_{st} 、姿态结构嵌入 p_{st} 和面具感知嵌入 i_{sm} ，这些嵌入共同输入扩散网络。

骨干 UNet 由交替的 ResNet 模块和 Transformer 模块组成，使模型能够同时捕捉局部纹理细节和全局上下文依赖关系。在训练过程中，目标是重构受到高斯噪声腐蚀的图像。优化目标是最小化预测噪声与真实噪声之间的重建误差，损失函数 $\mathcal{L}$ 定义为：

$$\mathcal{L} = \mathbb{E}_{x_0,\epsilon,t} \left\| \epsilon - \epsilon_\theta(x_t, f_{\text{st}}, p_{\text{st}}, i_{\text{sm}}, t) \right\|_2^2 . \tag{1}$$

为了进一步提高可控性和生成质量，我们在推理过程中采用分类器无关指导策略。通过加权和融合图像语义分支和姿势结构分支，有效地引导生成过程：

$$\hat{\epsilon}_\theta = w \cdot \epsilon_\theta(x_t, f_{\text{st}}, i_{\text{sm}}, t) + (1-w) \cdot \epsilon_\theta(x_t, p_{\text{st}}, t), \quad w \in [0,1]. \tag{2}$$

通过上述机制，PCDM 在驾驶员行为生成任务中展示了强大的姿态可控性和图像质量保证。生成的伪样本不仅在结构上与目标姿态对齐，而且在语义上保持一致，有效缓解了原始训练数据集的规模和分布的局限性。

A. 生成图像过滤机制

尽管 PCDMs 能够生成具有高度姿态一致性的图像，但语义上的歧义仍可能出现——例如动作分类错误（如发短信与打电话）或背景伪影。为了确保伪样本的语义纯净性，我们引入了大规模多模态模型 CogVLM 进行自动的图文一致性评估。

1) *行为类别和提示设计*：考虑到分心驾驶行为的多样性，我们根据数据集定义将动作空间分为 10 个不同的类别（C0–C9）。这些类别涵盖了典型的活动，如正常驾驶、用任一手打电话或发短信、调整多媒体系统、喝水、化妆、拿后座物品和与乘客交谈。

对于每个类别，我们构建一个简明准确的英语提示，引导 CogVLM 确定图像-文本对齐。这些提示是陈述性的且结构良好，旨在提升遵循指令的表现。表 I 列出了所有类别及其对应的提示。

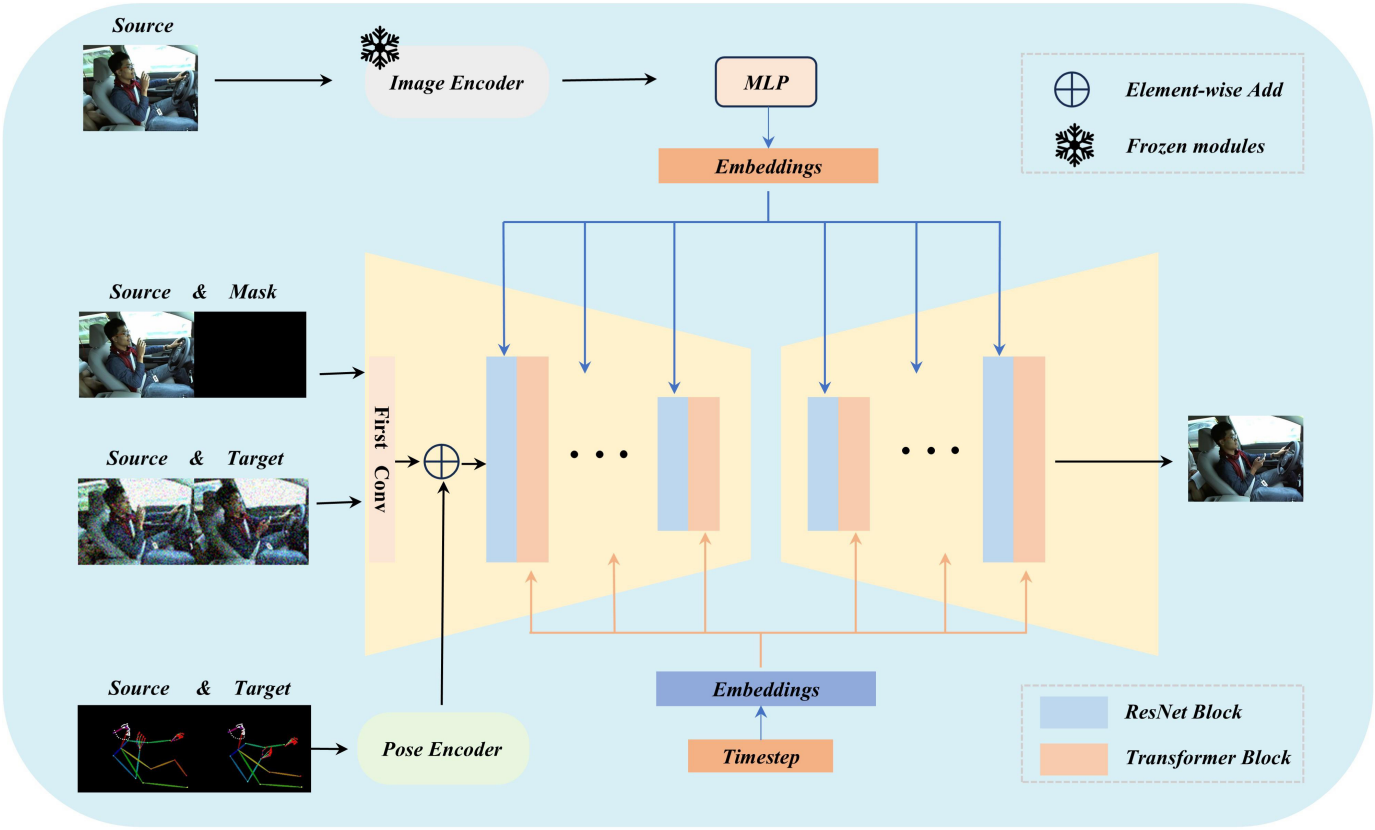

Fig. 1. 渐进条件扩散模型

2) 一致性评分函数：为了避免由离散的“是/否”输出引起的不可控性，本文设计了以下数值查询模板，强制 CogVLM 返回区间 $[0,1]$ 中的匹配评分：

> How well does this image match the description: “{ Prompt[c] } ” ?
> Respond with a number between 0 and 1, where 1 means perfect match.

图像-文本匹配函数定义为

$$s = C(I_g, \text{Prompt}[c]), \quad s \in [0,1],$$

其中 $C(\cdot)$ 表示在指令调优后的多模态相似性映射；I_g 是生成的伪样本；Prompt [c] 代表与类别 (C0,C1,···,C9) 对应的文本提示。s 越接近 1，语义一致性越高。

为了平衡数据纯度和保留率，本文设置了一个固定的阈值

$$\tau = 0.8.$$

筛选过程在算法 1 中展示。

最终的高质量伪样本集记为

$$D^{\star}_{\text{pseudo}} = \{I_g \mid s \geq \tau\}.$$

为了验证所提出的 PQ-DAF 框架在数据缺乏场景中的性能提升，我们在 PQ-DAF 中使用 ResNet50 [33] 作为分类模型，并在 StateFarm 和 AUC-DDD 数据集上在 10-shot 和 30-shot 设置下进行比较评估。

Algorithm 1 基于 CogVLM 的语义过滤

Require: Pseudo-sample set $\mathbb{S} = \{(I_g, c)\}$, prompt dictionary Prompt$[\cdot]$, threshold τ
Ensure: Filtered pseudo-sample set $\mathbb{S}_{\text{filtered}}$

```
S_filtered ← ∅
for all (I_g, c) in S do
    query ← “How well does this image match the description: ”Prompt[c]”? Respond with a number between 0 and 1, where 1 means perfect match.”
    s ← CogVLM(I_g, query)
    if s ≥ τ then
        S_filtered ← S_filtered ∪ {(I_g, c)}
    end if
end for
return S_filtered
```

B. 数据集

StateFarm 数据集是一个用于分心驾驶员检测的基准数据集，由 Kaggle 比赛发布。它包括 10 种分心类别（例如，用右手发短信、用左手打电话、操作收音机、正常驾驶），包含由仪表板摄像头拍摄的 17,462 张标记 RGB 图像。该数据集中涉及 26 名不同种族、肤色和性别的主体，确保了良好的多样性。对于少样本实验，我们使用原始训练集（80 % 的数据）对 PQ-DAF 的 PCDMS 模块进行微调，并随机抽取原始测试集（20 % 的数据）构建 10-shot 和 30-shot 子集（每类 10/30 张图像）作为少样本训练数据。

TABLE I
驾驶员分心行为类别及对应提示

Category	Behavior Description	Prompt
C0	Normal driving	The driver is driving normally with both hands on the steering wheel.
C1	Texting with right hand	The driver is texting with the right hand while driving.
C2	Holding phone to right ear	The driver is holding a phone to the right ear while driving.
C3	Texting with left hand	The driver is texting with the left hand while driving.
C4	Holding phone to left ear	The driver is holding a phone to the left ear while driving.
C5	Adjusting multimedia	The driver is adjusting the car's multimedia or infotainment system.
C6	Drinking water	The driver is drinking water while driving.
C7	Reaching toward back seat	The driver is reaching toward the back seat to grab something.
C8	Applying makeup	The driver is applying makeup while driving.
C9	Talking to passenger	The driver is talking to a passenger while driving.

AUC-DDD AUC-DDD（开罗美国大学分心驾驶员数据集）由开罗美国大学的 Abouelnaga 等人构建，广泛用于分心驾驶员检测。它包括 10 个类别（例如，安全驾驶、左手打电话、右手发短信、吃东西、打理、调节收音机），包含从 7 个国家的 31 名参与者在真实车辆中捕捉的 14,478 张图像，涵盖多种族群、肤色和性别。在小样本实验中，我们使用原始训练集（80 % 的数据）来微调 PQ-DAF 的 PCDMS 模块，并随机抽取原始测试集（20 % 的数据）来构建 10 张和 30 张样本子集（每类 10/30 张图像）作为小样本训练数据。

我们使用 Top-1 准确率作为评估指标来衡量模型性能，其定义为：其中 N 表示测试样本的总数量，y_i 是样本 i 的真实标签，$\hat{y}_i$ 表示模型预测的标签，$\mathbb{I}(\cdot)$ 是指示函数（如果预测正确则返回 1，否则返回 0）。

我们使用 F1-Score 作为附加的评估指标来衡量模型性能，定义为：

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

其中

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

TP 表示真正例，FP 表示假正例，而 FN 代表假负例。

C. 实施细节

所有实验均在配备 NVIDIA GeForce RTX 4070 Super GPU 和 Intel Core i5-13600KF CPU 的机器上进行。对于用于样本生成的 PCDMs，输入图像被统一调整为 512 × 512 。采用 Adam 优化器，学习率（lr）为 5×10^{-6} ，权重衰减为 1×10^{-3} 。训练进行了 30,000 个 epoch，批量大小为 5。为了减少内存使用，用于质量验证的 CogVLM 模型使用 4 位量化加载。

对于所有其他小样本学习实验，采用了 AdamW 优化器，其学习率为 1×10^{-4} ，权重衰减为 1×10^{-2} 。每个模型都进行了 20 个周期的训练，批量大小为 16，以确保实验设置的一致性和公平性。

TABLE II
在 StateFarm 数据集上的性能比较

Model	10-shot		30-shot	
	Top-1 (%)	F1 (%)	Top-1 (%)	F1 (%)
MobileNetv3$_{(ICCV'19)}$ [34]	16.67	16.17	21.33	20.71
MobileViT$_{(ICLR'22)}$ [35]	17.33	16.81	31.33	30.48
FastViT$_{(ICLR'22)}$ [36]	24.67	23.93	48.67	46.91
ConvNeXt $_{(CVPR'22)}$ [37]	25.33	24.57	42.67	41.12
Inceptionv4 $_{(AAAI'17)}$ [38]	26.67	25.87	60.00	58.24
PVT$_{(ICCV'21)}$ [39]	26.67	25.87	54.00	52.17
ResNet50$_{(CVPR'16)}$ [33]	36.67	35.57	64.67	62.38
Ours	54.00	52.38	88.00	85.44

D. 与最新方法的比较

1）在 StateFarm 上的对比：为了进一步验证所提出的伪样本增强方法在小样本驾驶行为识别任务中的有效性，我们在 StateFarm 数据集上进行了 10-shot 和 30-shot 分类实验，对应于现实场景中的极低和低样本条件。表格 II 展示了在原始训练集和增强集（aug）上的 Top-1 分类准确率的对比分析，涵盖了轻量级 CNN、标准 CNN 和基于 Transformer 的模型。

如表格 II 所示，我们提出的方法在 StateFarm 数据集上的表现显著优于所有比较模型，无论是在 10 次训练还是 30 次训练设置下，充分展示了其在数据稀缺条件下的有效性和泛化能力。具体而言，在 10 次训练设置中，我们的方法获得了 54.00 % 的 Top-1 准确率，比最好的基线模型 ResNet50（36.67 % ）提升了 17.33 个百分点。在 30 次训练设置中，准确率进一步提高到 88.00 %，远远超过了第二好的模型 Inceptionv4（60.00 %），同时也实现了 85.44 % 的 F1 分数，相比于 ResNet50 的 62.38 %。这一显著的改进可以归因于姿态引导图像生成策略与语义一致性过滤机制的协同效应，这不仅增加了训练样本的多样性，还保证了合成数据的高质量，使得模型在有限数据条件下能够学习到更为稳健和具辨识性的特征表示。

在验证了我们模型在 StateFarm 数据集上的效果之后，我们进一步在 AucDDD 数据集上评估我们的方法，以评估其在不同数据分布下的泛化能力。AucDDD 数据集包含更为多样的驾驶场景和行为模式，为分析我们的数据增强方法的有效性提供了补充的视角。

如表 ?? 所示，在更具挑战性的 AUC-DDD 数据集上，所提出的方法在不同的训练数据规模中均取得了最佳结果，显示出卓越的鲁棒性。在 10-shot 设置下，我们的方法达到了 40.67 % 的 Top-1 准确率，比表现最好的基线 ResNet50（30.67 % ）提高了 10.00 个百分点。在 30-shot 设置下，准确率进一步攀升至 71.33 %，远远超过第二好的模型 PVT（46.67 %），并实现了 71.27 % 的 F1 得分，这远高于 ResNet50 的 50.68 %。这些结果进一步验证了该方

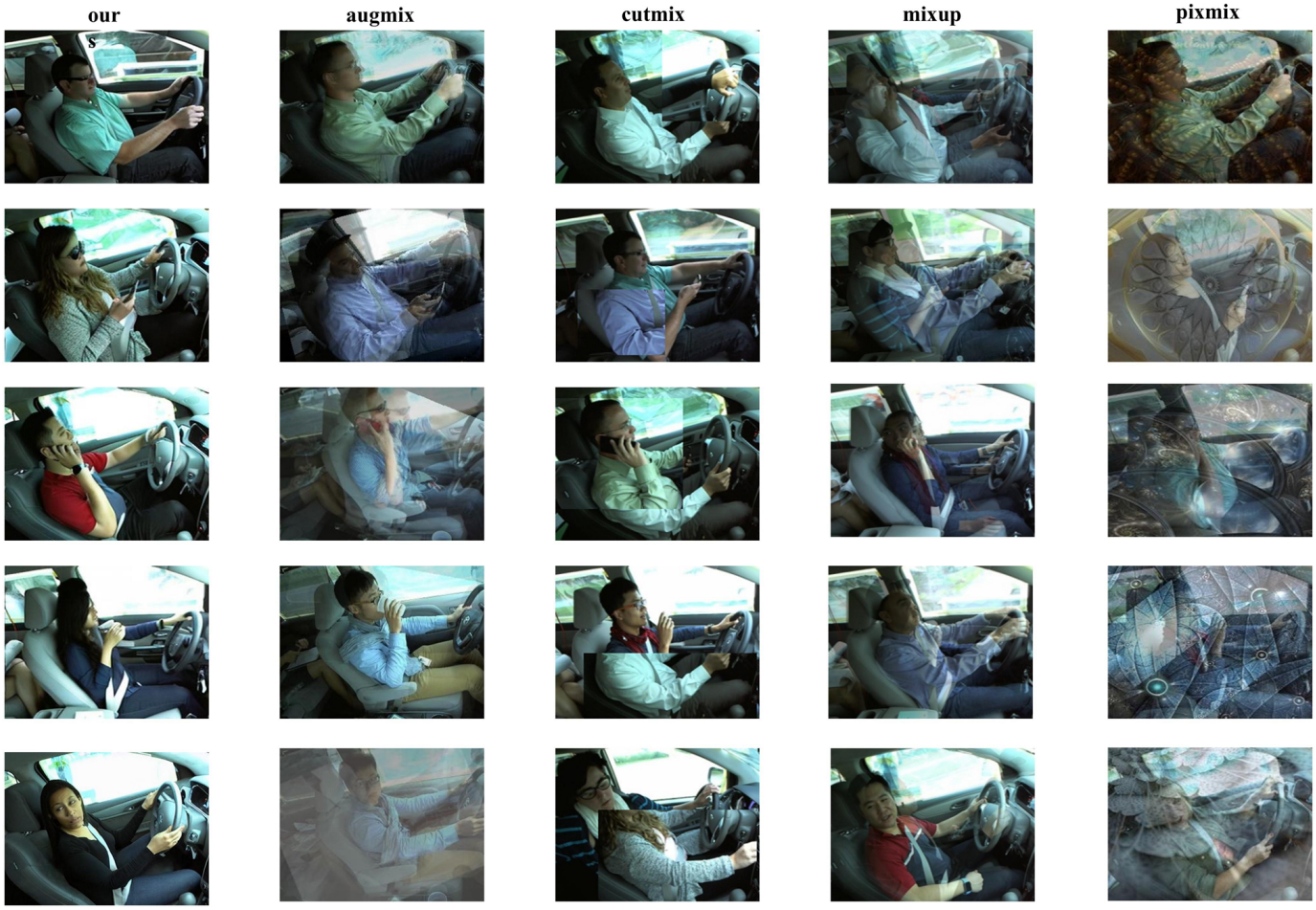

Fig. 2. 数据增强技术的视觉比较

TABLE III
不同数据增强方法的 Top-1 精度（%）比较

Method	Statefarm Top-1 Acc. (%)	AUC-DDD Top-1 Acc. (%)
AugMix [40]	40.67 ± 2.34	33.44 ± 1.82
Manifold [41]	39.46 ± 0.37	32.56 ± 1.60
Mixup [42]	39.67 ± 1.90	32.96 ± 1.73
CutMix [43]	39.78 ± 0.94	33.97 ± 1.45
PixMix [44]	33.44 ± 3.00	29.10 ± 1.25
PuzzleMix [45]	41.37 ± 0.82	34.42 ± 1.28
GuidedMixup [46]	42.45 ± 0.54	35.63 ± 1.17
Ours	54.67 ± 1.23	40.67 ± 1.12

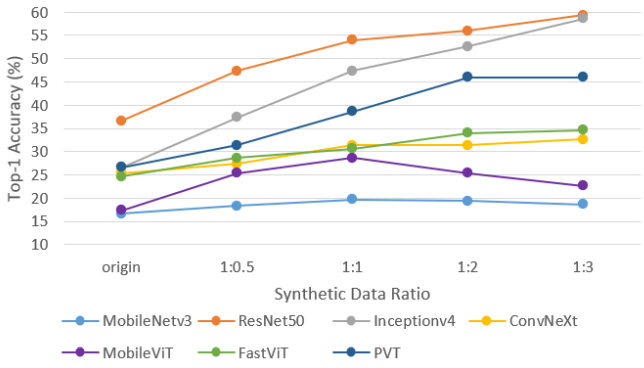

Fig. 3. 不同真实和合成数据混合比下的准确性比较

法在 AUC-DDD 数据集上的有效性，并且与在 StateFarm 数据集上观察到的性能趋势一致。

2) 数据增强比较：为了严格评估所提出的 PQ-DAF 在增强小样本驾驶员行为识别中的有效性，我们使用 ResNet-50 作为基准模型进行了对比评估。该评估包含了七种广泛采用的数据增强技术——AugMix、Manifold、Mixup、CutMix、PixMix、PuzzleMix 和 GuidedMixup，这些技术涵盖了多种策略，如像素级混合、语义引导混合和面向鲁棒性的转换。我们采用 Top-1 分类准确率作为主要指标，以量化比较各方法的性能。

表 III 表明，我们提出的数据增强方法在 StateFarm 和 AUC-DDD 数据集上显著优于几种现有的主流增强策略。例如，在 StateFarm 数据集上，我们的方法实现了 54.67 % 的 Top-1 精度，这比第二好的方法 GuidedMixup（42.45 %）高出超过 12 个百分点。在更具挑战性的 AUC-DDD 数据集上，我们的方法同样以 40.67 % 的 Top-1 精度领先，超越所有其他增强技术。值得注意的是，传统的合成

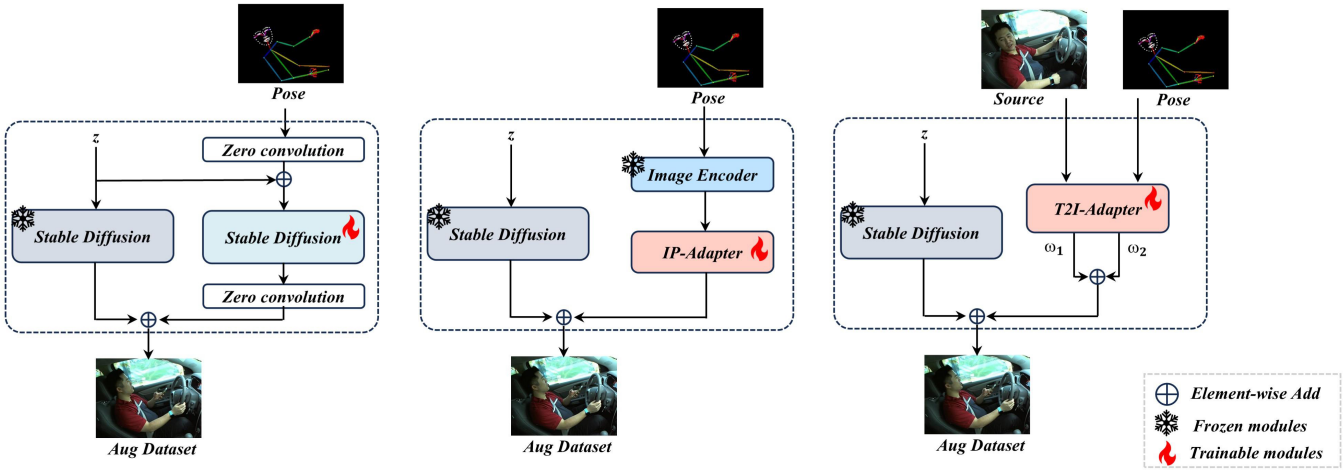

Fig. 4. PQ-ADF 框架的生成模型变体，其余过程完全相同

TABLE IV
在真实数据和合成数据不同混合比之下的准确性比较（10 次示例设定）

Model	Params (M)	Real Only	1:0.5	1:1	1:2	1:3
MobileViT (ICLR'22) [35]	0.95	17.33	25.33	28.67	25.33	22.67
MobileNetv3 (ICCV'19) [34]	1.53	16.67	17.33	18.67	19.33	18.67
FastViT (ICLR'22) [36]	3.24	24.67	28.67	30.67	34.00	34.67
PVT (ICCV'21) [39]	3.41	26.67	31.33	38.67	46.00	46.00
ConvNeXt (CVPR'22) [37]	4.83	25.33	27.33	31.33	31.33	32.67
ResNet50 (CVPR'16) [33]	25.56	36.67	47.33	54.00	56.00	59.33
Inceptionv4 (AAAI'17) [38]	41.10	26.67	37.33	47.33	52.67	58.67

Note : Real Only indicates using only real samples; 1:x indicates the ratio of real to generated data.

增强方法如 Mixup、CutMix 和 PuzzleMix 在两个数据集上的精度一般都在 30 % 左右，这表明有显著的性能差距。这些结果表明，我们的方法在提高模型判别力和泛化性能方面更有效。这可能归因于我们的方法生成的伪样本具有更高的语义一致性和上下文完整性，使模型能够在少样本学习条件下更好地利用潜在信息，从而实现更稳健的性能。

此外，我们提供了除了 PuzzleMix 和 GuidedMixup 之外所有方法生成的增强图像的可视化（因为这些方法是在潜在空间中进行的）。如图 2 所示，我们的方法在保存行为语义、清晰度和自然性方面实现了最佳表现。与 AugMix 和 Mixup 等方法相比，我们的增强样本保持了人体动作和场景结构的连贯性，避免了混合伪影和语义歧义。同时，与 CutMix 和 PixMix 产生的不自然遮挡或过于风格化的特征相比，我们的方法有助于模型更准确地捕捉关键行为线索。

3) *生成数据比例对模型性能的影响*：此实验旨在评估在极低数据场景下混合真实和合成数据的影响，其中用于训练的样本非常有限，并验证所提出的 PQE-AD 框架生成的合成数据在提升模型性能方面的有效性。具体而言，在每个类别 10 个样本的设置下，我们比较了 MobileNetv3、ResNet50、Inceptionv4、ConvNeXt、MobileViT、FastViT 和 PVT 这七个主流模型在不同真实与合成数据比例（1:0.5, 1:1, 1:2, 1:3）下的分类准确性。

如表 IV 和图 3 所示，在 10 次试验设置下，具有不同规模的模型在用不同比例的真实和合成数据训练时都表现出不同程度的性能提升，但提升的程度与模型容量密切相关。总体而言，随着合成数据比例的增加，大多数模型的分类准确性稳步提高，其中大容量模型受益最多。例如，ResNet50（25.56M 参数）在真实与合成比例为 1:0.5 时达到 47.33 % 的准确率，而在 1:3 时上升到 59.33 %——提高了 12 个百分点。同样，Inceptionv4（41.10M 参数）的准确率从 37.33 % 提高到 58.67 %，增加了超过 21 个百分点，展示了高容量模型在利用高质量合成样本方面的强大潜力。

相比之下，轻量级模型的增益更为有限，并且对混合比例更为敏感。例如，MobileNetv3（1.53M 参数）在 1:2 比例时达到其峰值性能，但提升幅度很小，而 MobileViT（0.95M 参数）在比例过高时甚至出现性能下降。这表明对于轻量级网络来说，适中的混合比例（建议在 1:1 到 1:2 之间）更为合适，因为它在保持样本多样性的同时保留了原始分布。过多的合成数据可能会引入噪音或领域偏差，由于模型容量有限，可能负面影响性能。因此，在实际应用中，建议根据具体模型调整比例：对于中到大容量的模型，逐步增加合成样本的比例可以获得更大的收益，而对于轻量级模型来说，采用 1:1 或 1:2 的保守比例更有利于稳定的性能提升。此实验验证了 PQE-AD 在极低数据条件下生成有用数据的有效性，并强调了在真实世界部署中不同容量模型的最佳混合策略差异。

E. 消融研究与分析

为了验证所提出的 PQ-ADF 框架在选择 PCDMs 模块方面的有效性，我们将不同的生成模型引入 PQ-ADF 框架中以构建三个变体，如图 4 所示。具体而言，PQ-ADF A 采用 ControlNet [47] 作为生成模型，使用提取的姿态信息作为条件输入来指导生成过程。虽然这种方法在一定程度上可以控制生成图像的姿态，但它不能约束身份特征，导致生成人物外观的随机变化。PQ-ADF B 使用 IP-Adapter [48] 作为生成模型，同样只使用姿态图像作为条件输入，因此存在无法控制人物特定特征的同样限制。PQ-ADF C 利用 T2I-Adapter [49] 作为生成模型。此方法支持灵活的多条件控制，在生成过程中，身份特征和姿态特征均作为条件输入提供，从而实现人物身份与姿态更精确的对齐。

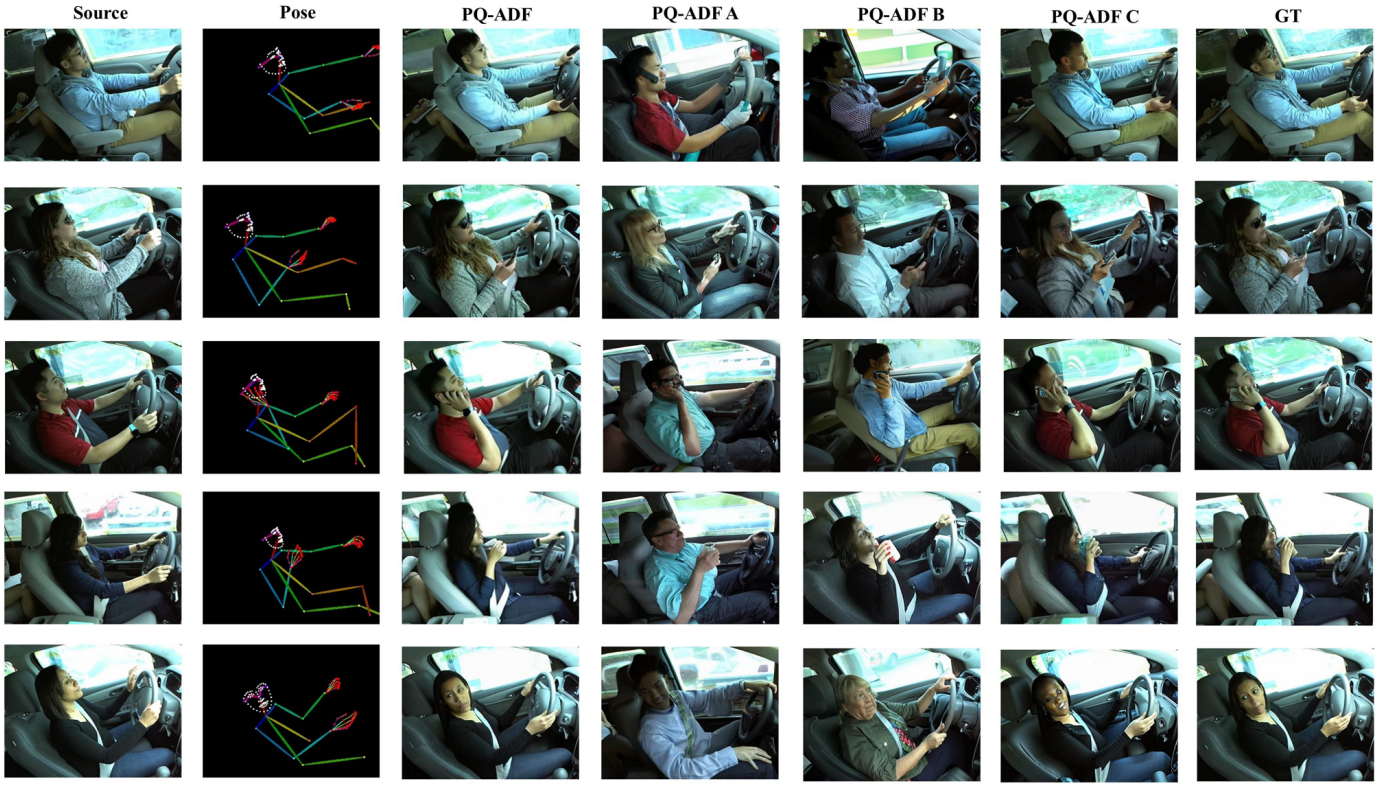

Fig. 5. 三个变体的视觉比较。

TABLE V
使用 ResNet-50 骨干网的生成模型准确性比较

Method	Top-1 Accuracy (%)
PQ-ADF A	45.67 ± 1.57
PQ-ADF B	43.78 ± 2.31
PQ-ADF C	49.92 ± 1.31
PQ-ADF	54.67 ± 1.23

Note : All models are evaluated on the same driver distraction dataset with 10-shot training.

如表 V 所示，使用所提出的 PCDMs 生成的样本取得了最佳的性能提升，准确率达到 54.67 %，显著优于其他生成方法。结果表明，PCDMs 模块通过多层次特征对齐，更适合于驾驶员分心检测任务，生成的驾驶员样本在外观和动作上与真实数据高度一致。

此外，我们还提供了各个变体生成的样本的视觉比较。如 5 所示，我们模型生成的样本在视觉上与真实数据更为一致。PQ-ADF C 在真实感方面排名第二，而 PQ-ADF A 和 PQ-ADF B 由于依赖于仅姿态条件，导致身份不一致和样本不够真实，从而产生较低质量的输出。

III. 结论

本研究提出了一种基于姿势驱动的质量控制数据增强框架（PQ-DAF），用于解决由于注释数量有限和域偏移导致的少样本驾驶员分心检测性能下降的问题。该框架采用了一种渐进条件扩散模型来合成姿势一致且多样化的样本，并利用视觉-语言模型进行语义质量过滤，从而有效增强了数据稀缺情况下的泛化能力。实验结果显示性能有显著提升，包括 ResNet50 的 Top-1 准确率从 36.67 % 提升到 54.00 %（在 StateFarm 数据集上），以及 Inceptionv4 从 19.33 % 提升到 34.00 %（在 AUC-DDD 数据集上）。尽管如此，该方法仍然依赖于姿势估计的准确性和过滤的鲁棒性。未来的工作将探索先进的姿势估计和多模态场景理解，以进一步提高语义保真度，优化样本选择，并扩展应用范围到更复杂的驾驶场景和行为类别。

References

[1] Z. Zhu, S. Wang, S. Gu, Y. Li, J. Li, L. Shuai, and G. Qi, “Driver distraction detection based on lightweight networks and tiny object detection,” Mathematical biosciences and engineering, vol. 20, no. 10, pp. 18 248–18 266, 2023.

[2] J. Lei, Z. Ni, Z. Peng, H. Hu, J. Hong, X. Fang, C. Yi, C. Ren, and M. A. Wasaye, “An intelligent network framework for driver distraction monitoring based on res-se-cnn,” Scientific Reports, vol. 15, no. 1, p. 6916, 2025.

[3] J. Cronje and A. P. Engelbrecht, “Training convolutional neural networks with class based data augmentation for detecting distracted drivers,” in Proceedings of the 9th International Conference on Computer and Automation Engineering, 2017, pp. 126–130.

[4] L. Zhu, Z. Zhang, D. Song, and B. Chen, “Analysis of traffic accident causes based on data augmentation and ensemble learning with high-dimensional small-

sample data," Expert Systems with Applications, vol. 237, p. 121782, 2024.
[5] I. K. K. Ugli, A. Hussain, B. S. Kim, S. Aich, and H.-C. Kim, "A transfer learning approach for identification of distracted driving," in 2022 24th International Conference on Advanced Communication Technology (ICACT). IEEE, 2022, pp. 420–423.
[6] J. Wang, Z. Wu, F. Li, and J. Zhang, "A data augmentation approach to distracted driving detection," Future internet, vol. 13, no. 1, p. 1, 2020.
[7] M. Z. Hasan, J. Chen, J. Wang, M. S. Rahman, A. Joshi, S. Velipasalar, C. Hegde, A. Sharma, and S. Sarkar, "Vision-language models can identify distracted driver behavior from naturalistic videos," IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 9, pp. 11 602–11 616, 2024.
[8] F. Shen, H. Ye, J. Zhang, C. Wang, X. Han, and W. Yang, "Advancing pose-guided image synthesis with progressive conditional diffusion models," arXiv preprint arXiv:2310.06313, 2023.
[9] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, S. XiXuan et al., "Cogvlm: Visual expert for pretrained language models," Advances in Neural Information Processing Systems, vol. 37, pp. 121 475–121 499, 2024.
[10] R. Tian, L. Li, M. Chen, Y. Chen, and G. J. Witt, "Studying the effects of driver distraction and traffic density on the probability of crash and near-crash events in naturalistic driving environment," IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 3, pp. 1547–1555, 2013.
[11] Y.-S. Poon, C.-Y. Kao, Y.-K. Wang, C.-C. Hsiao, M.-Y. Hung, Y.-C. Wang, and C.-P. Fan, "Driver distracted behavior detection technology with yolo-based deep learning networks," in 2021 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-ASIA). IEEE, 2021, pp. 01–05.
[12] A. Khellal, M. Boulahmar, A. Bahi, and A. Nemra, "Distracted driver detection using convolutional neural networks based segmentation model," in 2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC). IEEE, 2024, pp. 1–6.
[13] B. Qin, J. Qian, Y. Xin, B. Liu, and Y. Dong, "Distracted driver detection based on a cnn with decreasing filter size," IEEE transactions on intelligent transportation systems, vol. 23, no. 7, pp. 6922–6933, 2021.
[14] H. Mittal and B. Verma, "Cat-capsnet: A convolutional and attention based capsule network to detect the driver' s distraction," IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 9, pp. 9561–9570, 2023.
[15] C. Carney, D. McGehee, K. Harland, M. Weiss, and M. Raby, "Using naturalistic driving data to assess the prevalence of environmental factors and driver behaviors in teen driver crashes," 2015.
[16] Z. Fang, J. Chen, J. Wang, Z. Wang, N. Liu, and G. Yin, "Driver distraction behavior detection using a vision transformer model based on transfer learning strategy," in 2022 6th CAA International Conference on Vehicular Control and Intelligence (CVCI). IEEE, 2022, pp. 1–6.
[17] Z. Li, X. Zhao, F. Wu, D. Chen, and C. Wang, "A lightweight and efficient distracted driver detection model fusing convolutional neural network and vision transformer," IEEE Transactions on Intelligent Transportation Systems, 2024.
[18] W. Yang, C. Tan, Y. Chen, H. Xia, X. Tang, Y. Cao, W. Zhou, L. Lin, and G. Dai, "Birswint: Bilinear full-scale residual swin-transformer for fine-grained driver behavior recognition," Journal of the Franklin Institute, vol. 360, no. 2, pp. 1166–1183, 2023.
[19] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014.
[20] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable gans for pose-based human image generation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3408–3416.
[21] J. Gu and J. C. Ye, "Adain-based tunable cyclegan for efficient unsupervised low-dose ct denoising," IEEE Transactions on Computational Imaging, vol. 7, pp. 73–85, 2021.
[22] F. Shen, C. Wang, J. Gao, Q. Guo, J. Dang, J. Tang, and T.-S. Chua, "Long-term talkingface generation via motion-prior conditional diffusion model," arXiv preprint arXiv:2502.09533, 2025.
[23] F. Shen, J. Yu, C. Wang, X. Jiang, X. Du, and J. Tang, "Imaggarment-1: Fine-grained garment generation for controllable fashion design," arXiv preprint arXiv:2504.13176, 2025.
[24] F. Shen, X. Du, Y. Gao, J. Yu, Y. Cao, X. Lei, and J. Tang, "Imagharmony: Controllable image editing with consistent object quantity and layout," arXiv preprint arXiv:2506.01949, 2025.
[25] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, J. Laaksonen, M. Shah, and F. S. Khan, "Person image synthesis via denoising diffusion model," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 5968–5976.
[26] A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, "Multimodal garment designer: Human-centric latent diffusion models for fashion image editing," in Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 23 393–23 402.
[27] F. Shen, H. Ye, J. Zhang, C. Wang, X. Han, and Y. Wei, "Advancing pose-guided image synthesis with progressive conditional diffusion models," in The Twelfth International Conference on Learning Representations, 2024.

[Online]. Available: https://H.Bopenreview.H.Bnet/H.Bforum?H.Bid=rHzapPnCgTH.B

[28] X. Ju, A. Zeng, C. Zhao, J. Wang, L. Zhang, and Q. Xu, "Humansd: A native skeleton-guided diffusion model for human image generation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15 988–15 998.

[29] F. Shen and J. Tang, "Imagpose: A unified conditional framework for pose-guided person generation," Advances in neural information processing systems, vol. 37, pp. 6246–6266, 2024.

[30] F. Shen, X. Jiang, X. He, H. Ye, C. Wang, X. Du, Z. Li, and J. Tang, "Imagdressing-v1: Customizable virtual dressing," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 7, 2025, pp. 6795–6804.

[31] F. Shen, H. Ye, S. Liu, J. Zhang, C. Wang, X. Han, and Y. Wei, "Boosting consistency in story visualization with rich-contextual conditional diffusion models," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 7, 2025, pp. 6785–6794.

[32] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby et al., "Dinov2: Learning robust visual features without supervision," arXiv preprint arXiv:2304.07193, 2023.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[34] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan et al., "Searching for mobilenetv3," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1314–1324.

[35] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," arXiv preprint arXiv:2110.02178, 2021.

[36] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "Fastvit: A fast hybrid vision transformer using structural reparameterization," in Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 5785–5795.

[37] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11 976–11 986.

[38] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Proceedings of the AAAI conference on artificial intelligence, vol. 31, no. 1, 2017.

[39] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 568–578.

[40] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," arXiv preprint arXiv:1912.02781, 2019.

[41] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in International conference on machine learning. PMLR, 2019, pp. 6438–6447.

[42] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," arXiv preprint arXiv:1710.09412, 2017.

[43] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6023–6032.

[44] D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt, "Pixmix: Dreamlike pictures comprehensively improve safety measures," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16 783–16 792.

[45] J.-H. Kim, W. Choo, and H. O. Song, "Puzzle mix: Exploiting saliency and local statistics for optimal mixup," in International conference on machine learning. PMLR, 2020, pp. 5275–5285.

[46] M. Kang and S. Kim, "Guidedmixup: an efficient mixup strategy guided by saliency maps," in Proceedings of the AAAI conference on artificial intelligence, vol. 37, no. 1, 2023, pp. 1096–1104.

[47] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 3836–3847.

[48] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," arXiv preprint arXiv:2308.06721, 2023.

[49] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in Proceedings of the AAAI conference on artificial intelligence, vol. 38, no. 5, 2024, pp. 4296–4304.