arXiv:2603.29578v1 [cs.CV] 31 Mar 2026

用于面部表情识别的自适应间隔差异训练情绪扩散分类器

Rongkang Don g、Cuixin Yang、Cong Zhang、Yushen Zuo 和 Kin-Man Lam *IEEE* 高级会员 ,

摘要—面部表情识别（FER）对于人机交互至关重要，因为它使机器能够通过面部情感行为解读人类情绪和内部状态。尽管深度学习显著提升了 FER 的性能，但现有的大多数基于深度学习的 FER 方法都严重依赖判别式分类器来快速预测。这些模型往往会学到捷径，并且容易受到轻微分布偏移的影响。为解决这一问题，我们采用条件生成扩散模型，并提出用于 FER 的情绪扩散分类器（EmoDC），以增强对抗鲁棒性。然而，使用标准策略重新训练 EmoDC 无法惩罚错误的类别描述，导致识别性能欠佳。为改进 EmoDC，我们提出基于间隔的差异训练：在以正确类别描述为条件时鼓励准确预测，并惩罚以不匹配描述为条件的预测。该方法在正确与错误类别的噪声预测误差之间施加最小间隔，从而增强模型的判别能力。然而，固定间隔无法考虑不同图像在噪声预测方面的难度差异，限制了其效果。为克服这一局限，我们提出自适应间隔差异训练（AMDiT），针对每个样本动态调整间隔。大量实验表明，在 100 步评估中，与采用标准去噪扩散训练的基线模型相比，AMDiT 在 RAF-DB 基本子集、RAF-DB 复合子集、SFEW-2.0 数据集和 AffectNet 数据集上的准确率分别提升了 6.45%、16.29%、8.03% 和 4.86%。此外，在抵御噪声和模糊方面，EmoDC 的表现优于最先进的判别式分类器。

索引词—面部表情识别、扩散模型、扩散分类器。

I. 引言

面部表情是传达情绪状态的主要机制，对人类的非言语交流至关重要，
语言交流。自动面部表情识别（FER）使机器能够在人机交互过程中解读人类情绪并作出回应。现代深度学习技术取得的卓越成功显著提升了面部表情识别的性能 [1], [2]。当代面部表情识别深度学习方法主要依赖

Rongkang Dong、Cuixin Yang、Yushen Zuo 和 Kin-Man Lam 均任职于香港理工大学电气与电子工程系，地址为香港九龙（电子邮箱：rongkang97.dong@connect.polyu.hk、cuixin.yang@connect.polyu.hk、zuoyushen12@gmail.com、enkmlam@polyu.edu.hk）。

Cong Zhang 任职于山东大学控制科学与工程学院（电子邮箱：congzhang@sdu.edu.cn）

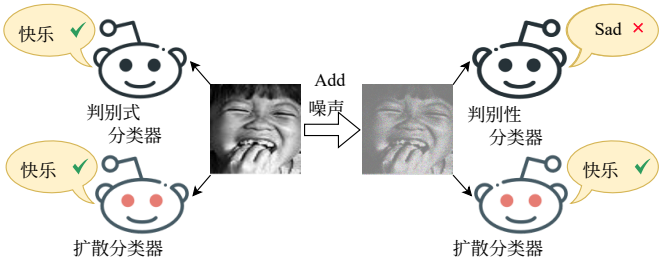


图 1：判别式分类器与扩散分类器在无噪声图像和噪声图像上的分类结果示例对比。

关于判别式分类器 [3]–[7]。给定输入样本$\boldsymbol{x}$，判别式分类器使用 softmax 函数输出面部表情类别上的类别概率分布 $p(\boldsymbol{c}|\boldsymbol{x})$，其中 $\boldsymbol{c}$ 表示类别标签。虽然这些模型能够高效地识别图像，但它们容易产生过度自信的预测 [8],、进行捷径学习 [9]–[11],，并且易受对抗攻击 [12]（示例见图 1）。最近，Li 等人 [11] 证明，判别式分类器往往会过度依赖伪相关特征，而非核心的任务相关特征，从而学习捷径，即使面对轻微的分布偏移也会失效。

尽管判别方法在面部表情识别中占主导地位，生成式分类器的潜力却仍未得到充分探索，这促使我们研究其在面部表情识别中的适用性。与容易出现捷径学习的判别模型不同，条件生成模型通过生成完整图像来避免这一问题，从而同时考虑核心特征和伪相关特征 [11]。因此，这些模型本质上减少了对非必要特征的依赖。条件生成模型专注于对似然 $p(\boldsymbol{x}|\boldsymbol{c})$进行建模，其中 $\boldsymbol{x}$ 和 $\boldsymbol{c}$ 分别表示样本及其类别描述。对于分类任务，使用贝叶斯定理计算 $p(\boldsymbol{c}|\boldsymbol{x})$ [13]:$p(\boldsymbol{c}|\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{c})p(\boldsymbol{c})/p(\boldsymbol{x})$。扩散模型 [14]–[16]，是一类最先进的生成模型，已在广泛的任务中展现出强大的生成能力 [15], [17]–[20]。条件扩散模型（例如 Stable Diffusion） [15], 会引入辅助文本输入以引导生成。近期研究已将这些模型改造为分类器，称为扩散分类器；它们展现出强大的零样本分类能力 [21],[22]以及固有的对抗鲁棒性 [23], [24]。扩散分

文献 [21] 中提出的融合分类器假设采用均匀先验$p(\boldsymbol{c})$，因此后验概率与似然成正比： $p(\boldsymbol{c}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{c})$。该方法根据类别描述计算噪声预测误差，并针对每个类别估计跨时间步和噪声的期望。这些期望与 $p(\boldsymbol{x}|\boldsymbol{c})$成反比，分类结果对应于期望误差最低的类别。

尽管扩散分类器展现出令人瞩目的零样本分类性能，但其零样本面部表情识别表现不佳，这一问题仍未得到解决，因为 Stable Diffusion 是基于图文对训练的，并非专为面部行为分析设计。更重要的是，扩散分类器需要计算所有时间步和类别的噪声预测误差，导致推理时间过长，难以在现实场景中部署。即使采用逐步剔除不可能类别，或在同一时间步共享噪声估计等注重效率的技术，[21], [22],推理速度慢仍然是一个持续存在的挑战。例如，尽管 100 步评估能够提高准确率，但其速度几乎比单步评估慢 200 倍。准确率和推理速度方面的这些局限表明，基于扩散的面部表情识别分类器仍需进一步改进。

提升扩散分类器面部表情识别准确率的一种直接方法，是在面部表情数据集上对其进行微调，由此发展出情绪扩散分类器（EmoDC）。为实现高效微调，采用低秩适配（LoRA） [25] ，显著减少了可训练参数的数量。EmoDC 的基本训练策略是，在正确的类别描述（正向提示） [14], [15]条件下，将噪声预测误差降至最低。然而，这一目标忽略了提高错误类别描述（负向提示）条件下噪声预测误差的重要性，限制了 EmoDC 区分不同面部表情的能力。此外，单步评估与多步评估之间仍存在显著的准确率差距。为解决这些局限，我们引入差异训练，在每个时间步上降低正向提示对应的噪声预测误差，同时放大负向提示对应的噪声预测误差。为进一步提升 EmoDC 的判别能力，我们提出基于间隔的差异训练，确保正向提示与负向提示对应的噪声预测误差之间至少存在一定差距。这不仅使模型能够区分面部表情，还能以明确的间隔进行区分。然而，使用固定间隔无法考虑不同输入之间噪声预测难度的差异。对于噪声预测容易的输入（即噪声预测误差较低），应采用较小的间隔；对于噪声预测困难的输入（即噪声预测误差较高），则需要较大的间隔。为此，我们提出自适应间隔差异训练（AMDiT），根据样本级难度动态调整间隔。具体而言，AMDiT 将非负描述条件下的噪声预测误差用作自适应间隔，使其能够反映不同噪声图像之间噪声预测难度的差异。大量实验验证了 AMDiT 在提升面部表情识别准确率以及显著缩小单步评估与多步评估之间性能差距方面的有效性。通过在每个时间步实现更准确的识别，AMDiT 能够以更少的时间步实现更快的推理，且准确率不会显著下降。此外，AMDiT 对含噪和模糊的面部图像表现出更强的鲁棒性，在具有挑战性的条件下优于最先进的判别式分类器。

本文的主要贡献总结如下:

- 我们开发了一种情绪扩散分类器（EmoDC），采用生成式扩散模型进行面部表情识别。据我们所知，这是首项探索将扩散分类器用于面部表情识别的工作。
- 我们提出了一种新颖的自适应边界差异训练方法，该方法根据样本级噪声预测难度，动态调整正向与负向提示之间的噪声预测误差间隔，从而有效增强 EmoDC 的判别能力并提升其面部表情识别性能。
- 大量实验表明，AMDiT 能够有效提升准确率，显著缩小单步评估与多步评估之间的准确率差距，并增强 EmoDC 对噪声图像和模糊图像的鲁棒性。

本文其余部分的组织如下。第二节回顾相关工作。第三节简要介绍预备知识，并详细阐述所提出的 AMDiT。第四节介绍实验设置并讨论实验结果。第五节对本文工作进行总结。

II. 相关工作

A. 面部表情识别的深度学习

数十年来，自动面部情感分析一直是一个活跃的研究课题。近年来，深度学习取得的进展显著推动了面部表情识别的发展，并实现了最先进的性能。基于卷积神经网络的面部表情识别方法采用卷积神经网络（CNN）[7], [26]–[28]提取空间特征，以编码面部表情信息 [29], [30]。例如，Kim 等人[27] 提出了 CVGG-19 网络，这是 VGG [26], 的改进变体，在显著提升性能的同时降低了计算成本。在 [3], [31],中，研究人员使用图来表示人脸，并结合卷积神经网络和图卷积网络提升面部表情识别性能。注意力机制 [32], [33], 也得到了应用，包括通道注意力、空间注意力和时间注意力，使模型能够分别聚焦于显著的通道、区域和帧。基于 Transformer 的面部表情识别方法，如 TransFER [34]和 VTFF [35], ，利用视觉 Transformer（ViT） [36]架构，通过自注意力机制提取具有全局交互性的面部特征。鉴于 ViT 的计算开销较高，APViT [37] 引入了一种注意力池化机制，用于丢弃信息价值较低的图像块和词元。该机制有效降低了计算成本，同时使模型能够聚焦于信息量更丰富的图像块和词元。基于多模态的面部表情识别方法利用多种模态，例如语音和文本描述。

图文表情识别模型 [6], [38], [39]通常依赖 CLIP [40],，后者利用自然语言监督进行视觉特征提取。此外，研究人员还整合了音频信息，以增强情绪识别 [41], [42]。虽然大多数研究主要关注基本面部表情，但人类往往会通过单一表情同时展现多种情绪。因此，复合表情识别 [4]，[5]，[43] 旨在识别更复杂的面部表情，例如惊喜以及愤怒与厌恶。尽管面部表情识别已取得长足进展，既有研究仍主要聚焦于判别式分类器，而对生成式分类器的探索相对不足。此外，面对受分布偏移影响的样本时，判别式面部表情识别方法的性能会下降，这仍是一项持续存在的挑战。本研究通过探讨扩散分类器在面部表情识别中的应用来弥补这些不足，利用其生成特性，提升模型在偏移数据分布下推理的可靠性。

B. 扩散分类器

生成模型估计数据似然 $p(\boldsymbol{x}|\boldsymbol{c})$，并可通过朴素贝叶斯 [13] 转换为分类器，以预测类别概率$p(\boldsymbol{c}|\boldsymbol{x})$。生成式分类器通常比判别式分类器更鲁棒 [23], [24]，[44], [45]。扩散模型是一类强大的生成模型，已在各种生成任务中展现出卓越性能 [15], [17]–[19]。为增强其分类能力，Mukhopadhyay等人 [46] 从引导扩散模型[47] 的中间层提取特征，并将其输入分类头。他们比较了四种类型的分类头，其中基于注意力的分类头表现最佳。这种分类方法本质上与判别式分类器的方法相似，即模型输出类别概率 $p(\boldsymbol{c}|\boldsymbol{x})$。CARD[48] 将扩散过程应用于 logit 空间，以处理分类和回归任务。当前的条件扩散模型（如 Stable Diffusion [15],）能够根据文本输入生成图像。由于经过大量图文对训练，Stable Diffusion在自然图像上的零样本分类性能表现出色 [21]，[22]。Jaini等人 [49] 分析了 [22] 中提出的扩散分类器，并揭示其物体识别行为与人类十分相似，尤其表现出形状偏差。提出鲁棒扩散分类器 [23], [24] ，旨在增强模型抵御对抗攻击和输入扰动的能力。然而，Stable Diffusion主要使用并非针对面部行为分析而构建的图文对进行训练，因此其理解细微面部情感行为的能力天生有限，导致零样本表情识别性能不佳。为解决这一局限，我们引入 LoRA，在面部表情数据集上高效微调扩散模型，使模型具备面部表情分析所需的领域专属知识。为进一步提升性能，我们提出一种新颖的 AMDiT 方法，通过同时提高准确率、推理速度以及分布偏移下的鲁棒性，显著增强 EmoDC模型。

C. 三元组损失

三元组损失是机器学习[50], [51]中广泛使用的一种函数。它旨在学习一个嵌入空间，使同一类别中锚点样本与正样本之间的距离最小化，同时使锚点样本与不同类别中负样本之间的距离最大化。FaceNet [51]率先将三元组损失用于深度人脸识别，并提出一种三元组选择策略，以在每个小批量中挖掘困难样本。Wu 等人 [52] 引入距离加权间隔，为三元组损失训练选择信息量更丰富且更稳定的样本。为加快收敛，研究者提出了基于三元组学习的多种推广方法。多类别*N*对损失 [53] 为每个锚点样本引入多个负样本，而 Magnet 损失 [54]则将嵌入划分为类别特定簇，以减少局部分布之间的重叠。然而，迄今为止，三元组概念仅用于依赖判别式分类器的识别任务。基于这一概念，我们为生成式扩散分类器提出一种新颖的训练方法 AMDiT。AMDiT会降低正向文本描述的噪声预测误差，同时提高负向文本描述的噪声预测误差，并根据各样本对应的噪声预测难度动态调整其误差间隔。

III. 情绪扩散分类器

本节首先对 EmoDC 进行初步介绍，随后介绍四种训练方法，并通过这些方法逐步提出新颖的 AMDiT 方法。

A. 初步介绍

1）扩散模型：扩散模型 [14], [55]是一类生成模型，通过迭代地添加和移除噪声来学习生成数据。这涉及两个主要过程：前向过程和反向过程。给定来自未知真实分布 q（$\boldsymbol{x}$）的观测值$\boldsymbol{x}_0$ 以及扩散步数总数 T，前向过程按照预先设定的方差调度 $\{\beta_t\}_{t=1}^T$，逐步添加高斯噪声 $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$，以生成一系列含噪样本 $\{\boldsymbol{x}_t\}_{t=1}^T$。前向过程可表示如下：

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \tag{1}$$

其中 $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t;\ \sqrt{1-\beta_t}\boldsymbol{x}_{t-1},\ \beta_t\boldsymbol{I}\right)$。前向过程的一个性质是，可以按如下方式简单地获得任意时间步 t处的含噪图像 $\boldsymbol{x}_t$：

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\overline{\alpha}_t}\boldsymbol{x}_0, (1-\overline{\alpha}_t)\boldsymbol{I}\right), \tag{2}$$

其中 $\alpha_t = 1 - \beta_t$ 和 $\alpha_{\overline{t}} = \prod_{s=1}^{t} \alpha_s$。

反向过程通过学习到的高斯转移，逐渐从高斯分布 $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T; \mathbf{0}, \boldsymbol{I})$ 中去除噪声。该过程定义如下：

$$p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \cdot \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \tag{3}$$

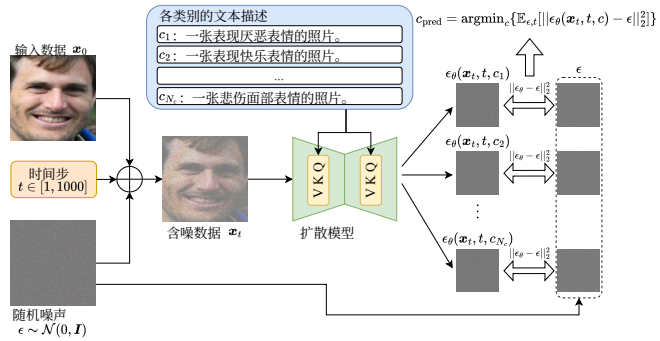


图2：用于面部表情识别的扩散分类器推理流程 [21] 。 N_c 表示类别总数。

其中 $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)\ \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t))$。目标是训练一个模型，使其通过参数 θ 学习函数 $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ 和$\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$。Ho 等人 [14] 推导出了一个去噪扩散模型，其目标是预测噪声。该模型通过优化对数似然的近似证据下界，具体如下：

$$\log p_\theta(\boldsymbol{x}_0) \geq -\mathbb{E}_{\epsilon,t}[w_t||\epsilon_\theta(\boldsymbol{x}_t,t)-\epsilon||_2^2] + A, \tag{4}$$

其中，$\epsilon_\theta(\boldsymbol{x}_t, t)$表示可训练模型在时间步 t 预测的噪声，w_t 是时间步 t的噪声预测误差权重，A 是一个常数。他们通过对 $t \in [1,1000]$设置 $w_t = 1$，简化了公式（4），得到：

$$-\mathbb{E}_{\epsilon,t}[||\epsilon_\theta(\boldsymbol{x}_t,t)-\epsilon||_2^2] + A. \tag{5}$$

扩散模型的标准训练目标记为$\mathcal{L}_{\text{base}}$，旨在最小化噪声预测误差，表示如下：

$$\mathcal{L}_{\text{base}} = \mathbb{E}_{\boldsymbol{x}_0,\epsilon,t}[||\epsilon_\theta(\boldsymbol{x}_t,t)-\epsilon||_2^2], \tag{6}$$

其中，常数 A被忽略，因为它不依赖于模型参数 θ。

2）扩散分类器：本节简要介绍扩散分类器 [21]。扩散分类器采用条件扩散模型，例如 Stable Diffusion [15]，，其中类别文本描述作为条件输入。基于这一条件输入，公式（4）和公式（5）可改写如下 [21]:

$$\log p_\theta(\boldsymbol{x}_0|c) \geq -\mathbb{E}_{\epsilon,t}[||\epsilon_\theta(\boldsymbol{x}_t,t,c)-\epsilon||_2^2] + A, \tag{7}$$

其中，c表示类别条件，即图像的文本描述。

图2展示了扩散分类器的分类过程。模型以所有类别描述为条件，针对时间步 t下的输入图像 $\boldsymbol{x}_0$ 生成预测噪声。通过评估多个时间步和噪声，得出分类结果如下：

$$c_{\text{pred}} = \text{argmin}_c\{\mathbb{E}_{\epsilon,t}[||\epsilon_\theta(\boldsymbol{x}_t,t,c)-\epsilon||_2^2]\}. \tag{8}$$

先前研究 [21], [22] 指出，扩散分类器具有出色的零样本分类性能：对于正确的类别描述，其噪声预测误差通常较低；对于错误的类别描述，误差则较高。然而，推理速度明显较慢，因为必须针对（a）每个类别描述、（b）每个时间步以及（c）大量采样噪声计算噪声预测（如公式（8）所示）。为提高效率，研究采用了候选类别剪枝方法 [21], [22] ，通过对少量时间步上的误差取平均，逐步剔除不太可能的类别。此外， [22],采用了共享噪声，即在给定时间步 t，对所有类别评估同一个噪声样本。在我们的实验中，我们采用共享噪声方法来提高推理效率。

B. EmoDC

扩散分类器在自然图像上展现出强大的零样本性能。然而，其零样本表情识别性能较差，准确率低于 27%（见表III）。这一局限源于训练过程中面部情感分析数据不足。

在本研究中，我们将扩散分类器用于面部表情识别，并将其称为情绪扩散分类器（EmoDC）。为使模型具备面部表情知识，我们使用 LoRA [25],这一高效训练方法对扩散模型进行微调。然而，Stable Diffusion [15] 仅使用正图文对进行训练，并未显式增大负向配对的噪声预测误差，这限制了模型的判别能力。为解决这一问题，我们提出增大负图文对的噪声预测误差。

1）对比噪声差异训练（*CoDiT*）：一种简单直观的实现方式是采用 CoDiT，其目标函数定义如下：

$$\mathcal{L}_{\text{CoDiT}} = \mathbb{E}_{\boldsymbol{x}_0,t,\epsilon}\left[||\epsilon_\theta(\boldsymbol{x}_t,t,c_p)-\epsilon||_2^2 - \lambda_1||\epsilon_\theta(\boldsymbol{x}_t,t,c_n)-\epsilon||_2^2\right], \tag{9}$$

其中，c_p 和 c_n 分别表示正向提示和负向提示，λ_1 表示与负向图文对相关误差的权重。负向提示 c_n 是错误的文本描述，例如，对于一张展示快乐面部表情的图像，使用“A photo of a disgusted facial expression”这样的描述（如图3(a)所示）。在实验中，c_n 是随机选取的。CoDiT 明确地最大化正向配对与负向配对之间的噪声预测误差差异。尽管基于间隔的差异损失有利于分类，但调整超参数λ_1时会遇到挑战。一方面，λ_1取值较大时（如 0.5 或 1），训练过程会变得不稳定——随着训练的进行，正向配对和负向配对的预测误差都会大幅增加，尽管二者的差异也随之增大。这与扩散模型旨在最小化正向描述误差的目标相悖，并削弱了模型估计真实数据分布的能力。另一方面，λ_1取值较小时（如 0.01 或 0.005），负向图文对的贡献会受到限制，从而降低了增加负向描述噪声预测误差的力度。因此，模型受到的区分正向配对与负向配对的激励较弱，限制了 EmoDC 性能的提升。

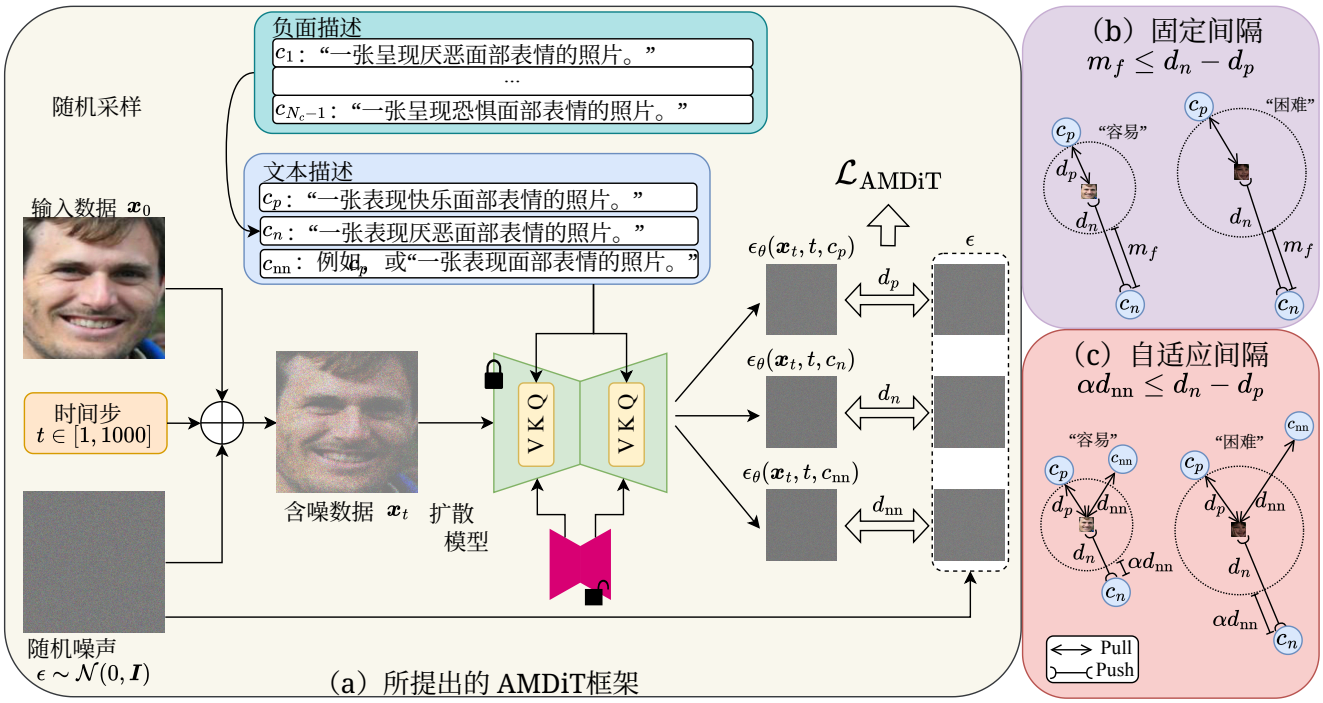


图3：示意图：(a) 所提出的 AMDiT框架，其中 d_* 表示噪声预测误差 $||\epsilon_\theta(\boldsymbol{x}_t, t, c_*) - \epsilon||_2^2$，* 表示 p、n 或 nn。具体而言，d_p、d_n 和 d_{nn} 分别表示正图文对、负图文对，以及由人脸图像和非负面文本提示组成的一对数据的噪声预测误差。这些预测误差用于计算 $\mathcal{L}_{\text{AMDiT}}$ 训练目标（式(11)）。图像的负向提示 c_n 从负面描述池中随机抽取（即除正面描述之外的所有类别描述）。(b) 使用固定间隔 m_f进行训练的分析。(c) 使用自适应间隔 αd_{nn}进行训练的分析。圆的半径越大，表示噪声预测难度越高，因此误差也越大（例如“困难”样本），反之亦然。

2) 基于间隔的差异训练：受三元组损失 [51], 启发，我们引入一种基于边界的差异损失来增强模型训练，具体定义如下：

$$\mathcal{L}_{\text{margin}} = \mathbb{E}_{\boldsymbol{x}_0,t,\epsilon}\left[\max\left\{||\epsilon_\theta(\boldsymbol{x}_t, t, c_p) - \epsilon||_2^2 - ||\epsilon_\theta(\boldsymbol{x}_t, t, c_n) - \epsilon||_2^2 + m, 0\right\}\right], \tag{10}$$

其中，m 表示间隔。

固定边界差异训练（*FMDiT*）：在这种训练策略中，间隔参数“m”通常在整个训练过程中保持固定（记为 m_f）。较大的固定间隔可能导致不稳定（类似于 CoDiT 中较大的 λ_1 所产生的影响），从而降低模型的噪声预测能力，并放大正样本对的误差。相反，较小的间隔虽能稳定训练，却无法考虑噪声预测难度的差异。例如，为容易样本分配较大的间隔可能会损害模型的噪声预测能力，因为过度增大负向提示上的噪声预测误差，也会导致正向样本上的噪声预测误差增加。另一方面，对困难样本使用较小的间隔，对模型判别能力的影响有限。因此，FMDiT 缺乏适应性，限制了其性能提升潜力。

自适应间隔差异训练（*AMDiT*）：为了解决使用固定间隔的局限性，我们提出了 AMDiT，这是一种旨在提升 EmoDC 性能的新型框架，如图3(a)所示为提出的 AMDiT流程。我们引入了以非负面文本提示 c_{nn} 为条件的噪声预测误差，并将其用作自适应间隔。该非负面描述可以是正向提示 c_p ($c_{\text{nn}} = c_p$)、不包含类别信息的文本描述（记作 c_{nc}，例如“面部表情的照片。”），或空文本（$c_{\text{nn}} = \emptyset$）。提出的 AMDiT 的目标函数定义如下：

$$\mathcal{L}_{\text{AMDiT}} = \mathbb{E}_{\boldsymbol{x}_0,t,\epsilon}\left[\max\left\{||\epsilon_\theta(\boldsymbol{x}_t, t, c_p) - \epsilon||_2^2 - ||\epsilon_\theta(\boldsymbol{x}_t, t, c_n) - \epsilon||_2^2 + \alpha||\epsilon_\theta(\boldsymbol{x}_t, t, c_{\text{nn}}) - \epsilon||_2^2, 0\right\}\right], \tag{11}$$

其中，α 控制自适应间隔的权重。我们将 α 设为较小值，以确保训练稳定性。

采用这种自适应间隔主要基于两个关键考量：

首先，固定间隔无法适应噪声预测难度各异的样本。图 3(b) 和图 3(c) 对比了固定间隔方案与自适应间隔方案下的训练过程。固定间隔方案对所有样本统一采用静态的预设间隔，忽略了噪声预测难度的差异。相比之下，自适应间隔方案会根据样本特定的难度动态调整间隔。

其次，去噪扩散模型 [14], [15] 经过训练，能够从含噪图像或潜在表示中预测噪声。

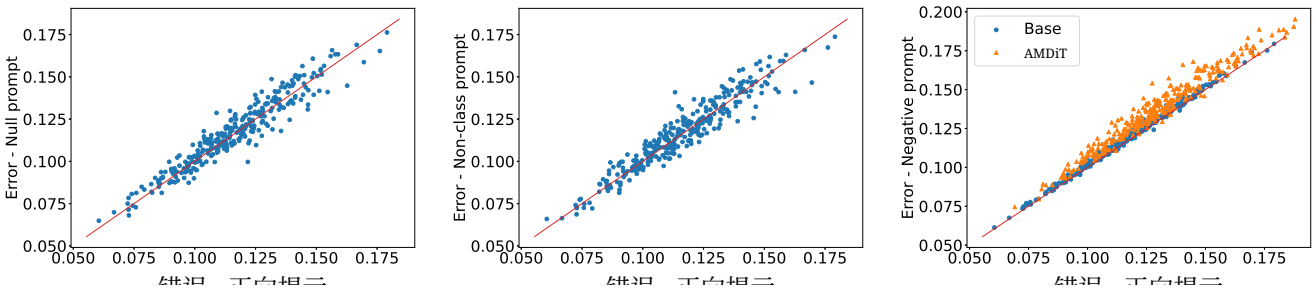


图 4：正向提示 c_p 与空提示 ∅、非类别提示 c_{nc} 和负向提示 c_n之间的噪声预测误差相关性。对于每个样本，负向提示的最终误差通过对所有错误类别描述的误差取平均值计算得出。红线表示斜率为 1 的线性参考线。“基线”模型是使用普通微调训练的 EmoDC，未采用负图文对，其训练目标定义于公式（6）中。实验在 RAF-DB B 的 300 个样本上进行。

虽然噪声预测性能会随不同样本的预测难度而变化，但对于同一图像，在不同非负描述条件下，预测误差的变化很小。图 4(a) 和图 4(b) 展示了正向提示与空提示、非类别提示所对应的噪声预测误差之间的相关性，为这一观察提供了支持。直观而言，噪声预测误差可以作为噪声预测难度的指标：误差越小，预测难度越低；误差越大，预测难度越高。图 4(a) 和图 4(b) 清楚地显示，不同非负提示（例如空提示与正向提示）之间存在斜率为 1 的强线性相关性。这表明，噪声预测难度主要由含噪潜变量（在 Stable Diffusion 中）的特性决定，文本提示的影响很小。因此，可以将非负提示对应的噪声预测误差用作自适应间隔，以反映样本特定的预测难度。这样一来，较简单的样本会被分配较小的间隔，较困难的样本则会被分配较大的间隔。图 4(c) 展示了正向提示与负向提示对应误差之间的相关性。值得注意的是，大多数数据点位于红色参考线（斜率为 1）上方，表明负向提示对应的误差始终高于正向提示。这种差异凸显了在面部表情识别任务中采用扩散模型的有效性。此外，与 EmoDC Base 模型相比，所提出的带有 AMDiT 的 EmoDC ($c_{\text{nn}} = c_p$) 在负向提示下表现出更高的误差，这表明 AMDiT 提升了扩散模型区分面部表情的判别能力。

IV. 实验

A. 数据集

为验证所提出方法的有效性，我们在四个真实场景下面部表情数据集上进行了评估，涵盖基本面部表情和复合面部表情：RAF-DB 基本集（记作 RAF-DB B）、RAF-DB 复合集（记作 RAF-DB C） [29], SFEW-2。0 [56], 以及 AffectNet [57]。此外，我们还采用 FER2013Plus[58]进行跨数据集评估。

RAF-DB [29] 是一个大规模数据集，包含近 30,000 个图像样本，分为两个子集：基本集和复合集。以往的大多数研究都聚焦于基本集。基本集包含 12,271 张训练图像和 3,068 张测试图像，分为七个类别：一种中性面部表情（Ne）和六种基本面部表情，即快乐（Ha）、愤怒（An）、悲伤（Sa）、厌恶（Di）、惊讶（Su）和恐惧（Fe）。另一方面，复合集包含 3,162 个训练样本和 792 个测试样本，涵盖十一种复合面部表情：HaDi（即快乐-厌恶）、HaSu、SaFe、FeAn、FeSu、SaAn、SaSu、SaDi、AnSu、AnDi 和 DiSu。采用这两个子集来展示所提出的方法在分析简单和复杂面部表情方面的有效性。

SFEW-2.0 [56] 是一个具有挑战性的人脸表情识别基准，包含从电影中提取的 1,500 多张面部图像。该数据集分为三部分：包含 958 个样本的训练集、包含 436 个样本的验证集，以及包含 372 个样本的测试集。我们遵循先前研究 [3], [37], [59] 的做法，使用验证集进行评估。SFEW 2.0 包含七个类别，与 RAF-DB B 中的类别相同。

AffectNet [57] 是一个大规模面部表情识别数据库，包含超过 290,000 张面部表情图像。在本研究中，我们采用七种标准面部表情类别。该数据集分为两个子集：包含 283,900 个样本的训练集和包含 3,500 个样本的验证集。验证集是平衡的，每个类别的样本数量相同。

FER2013Plus [58]是 FER2013[60] 的重新标注版本。该数据集分为三个子集：包含 28,558 个样本的训练集、包含 3,579 个样本的验证集，以及包含 3,573 个样本的测试集。我们采用测试集中的七种基本表情进行跨数据集评估。

表I：使用 CoDiT、FMDiT 和 AMDiT 分别调优的 EmoDC 变体，在不同 λ_1、m_f 和 α 超参数取值下于 RAF-DB B 上的准确率（%）。

方法	超参数	1步	10步	100 步	Avg.
Base	-	55.74	76.27	83.67	71.89
CoDiT	$\lambda_1 = 0.001$	56.71	77.71	84.45	72.96
	$\lambda_1 = 0.005$	58.47	77.97	85.17	**73.87**
	$\lambda_1 = 0.01$	56.71	77.90	84.45	73.02
FMDiT	$m_f = 0.0001$	78.85	86.44	86.83	84.04
	$m_f = 0.0005$	87.35	87.03	87.22	**87.20**
	$m_f = 0.001$	86.96	86.99	87.19	87.05
AMDiT	$\alpha = 0.005$	88.72	89.28	89.41	89.14
	$\alpha = 0.01$	89.18	89.83	90.12	**89.71**
	$\alpha = 0.05$	88.23	88.66	88.62	88.50

B. 实现细节

我们采用 Stable Diffusion v2.1[1] [15] 作为条件扩散模型，以评估所提出训练方法的性能。为提高训练效率，我们使用秩为 16 的 LoRA。优化器采用 AdamW [61]，学习率设为 0.0001，权重衰减设为 0.01。梯度裁剪的最大值设为 1，梯度累积步数为 1。RAF-DB B、RAF-DB C 和 SFEW 2.0 的训练步数分别设为 30,000、10,000 和 2,000。实验在 RTX4090 GPU 上进行，总批次大小为 64。模型使用分辨率为 256×256 的图像进行训练。使用 MTCNN [62], 对图像进行对齐和裁剪，分辨率设为 512×512，随后调整为 256×256。随机水平翻转是唯一采用的数据增强技术。

C. 实验结果

1）超参数研究：我们研究了不同超参数取值—— λ_1、m_f和 α ——分别对 CoDiT、FMDiT 和 AMDiT 的影响。EmoDC Base 作为基线方法，采用原始训练方法，不进行差异训练（公式 (6)）。评估采用不同数量的时间步，包括单步评估，以及均匀间隔的 10 步和 100 步评估。最终性能通过对单步、10 步和 100 步评估的准确率取平均值来衡量。单步评估中，Base 和 CoDiT 使用第 400 个时间步，而 FMDiT 和 AMDiT 使用第 100 个时间步（参见第四节 IV-C4 的分析）。本实验中，AMDiT 采用 $c_{\text{nn}} = c_p$。实验结果列于表I，表明模型性能会随超参数设置而变化。对于所有方法，这些超参数取值过大都可能导致训练不稳定和准确率欠佳。对于 CoDiT，λ_1 较小，无法充分强调负图文对，因此相较于 EmoDC Base 仅有微小提升。对于基于间隔的方法，较小的 m_f 或 α 会因分离间隔不足而削弱 EmoDC 的判别能力。通过实验确定的最佳超参数设置如下：CoDiT 为 $\lambda_1 = 0.005$，FMDiT 为 $m_f = 0.0005$，AMDiT 为 $\alpha = 0.01$。后续所有实验均采用这些设置。

[1]我们采用 Hugging Face 提供的 Stable Diffusion v2.1

表II：AMDiT 在 RAF-DB B 上使用不同非负向提示时的准确率（%）。此处，非类别提示为c_{nc} =“一张面部表情照片。”

$c_{\text{nn}} =$	c_{nc}	c_p	$\emptyset$
Acc.	89.08	**90.12**	87.74

表III：Stable Diffusion 使用不同训练方法进行 100 步评估时的准确率（%）。

方法 \ 数据集	RAF-DB B	RAF-DB C	SFEW-2.0	AffectNet
零样本	26.34	17.68	24.31	28.14
EmoDC Base	83.67	51.89	48.39	59.63
CoDiT	85.17	53.54	49.77	60.26
FMDiT	87.22	63.38	52.98	61.63
AMDiT ($c_{\text{nn}} = c_{\text{nc}}$)	89.08	63.51	**56.42**	**64.49**
AMDiT ($c_{\text{nn}} = c_p$)	**90.12**	**68.18**	56.19	63.89

2）非负提示分析：FMDiT 无法适应样本级噪声预测难度。为解决这一问题，我们提出 AMDiT，该方法使用根据非负提示条件下的噪声预测误差计算得出的自适应间隔。此误差反映了每个样本的噪声预测难度。为分析不同非负提示 c_{nn}的影响，我们在 RAF-DB B 上使用 c_{nn} =“一张面部表情照片。”（c_{nc}）、 $c_{\text{nn}} = c_p$和 $c_{\text{nn}} = \emptyset$开展实验。除第四节 IV-C4 中的实验外，本实验及后续实验均采用 100 步评估的结果。如表II所示，三种自适应间隔方法的表现均优于 Base、CoDiT 和 FMDiT（见表I）。这证明了所提出的 AMDiT 方法的有效性。在三种自适应间隔方法中，采用$c_{\text{nn}} = c_p$ 的 AMDiT 取得了最佳结果。

3) 增量分析：为验证所提出方法的有效性，我们在四个数据集上开展了实验。本评估包含六种方法，包括零样本分类、基线 EmoDC（Base），以及采用差异训练技术的四种增强变体：CoDiT、FMDiT、AMDiT ($c_{\text{nn}} = c_{\text{nc}}$) 和 AMDiT ($c_{\text{nn}} = c_p$)。零样本分类使用分辨率为 512×512 的图像进行评估，与其原始训练配置一致；其他方法则采用分辨率为256×256 的图像。

如表III所示，与零样本分类相比，在面部表情数据集上训练的方法显著提升了面部表情识别性能。基于差异的训练方法，包括 CoDiT、FMDiT 和 AMDiT，通过抑制负图文对的准确噪声预测，表现优于 EmoDC Base。具体而言，在 RAF-DB B、RAF-DB C、SFEW-2.0 和 AffectNet 上，CoDiT 相较于基线模型的准确率分别提升了 1.50%、1.65%、1.38% 和 0.63%。基于间隔的差异训练方法，即 FMDiT 和 AMDiT，则进一步提升了性能。例如，在 RAF-DB B 上，FMDiT 取得了显著的 3.55% 提升，准确率达到 87.22%。然而，FMDiT 难以根据噪声预测难度的不同程度调整间隔。相比之下，采用自适应间隔的 AMDiT 在所有数据集上均表现出色。具体而言，AMDiT ($c_{\text{nn}} = c_{\text{nc}}$) 变体相较于 EmoDC Base，在 RAF-DB B、RAF-DB C、SFEW-2.0 和 AffectNet 上的准确率分别提升了 5.41%、11.62%、8.03% 和 4.86%，相

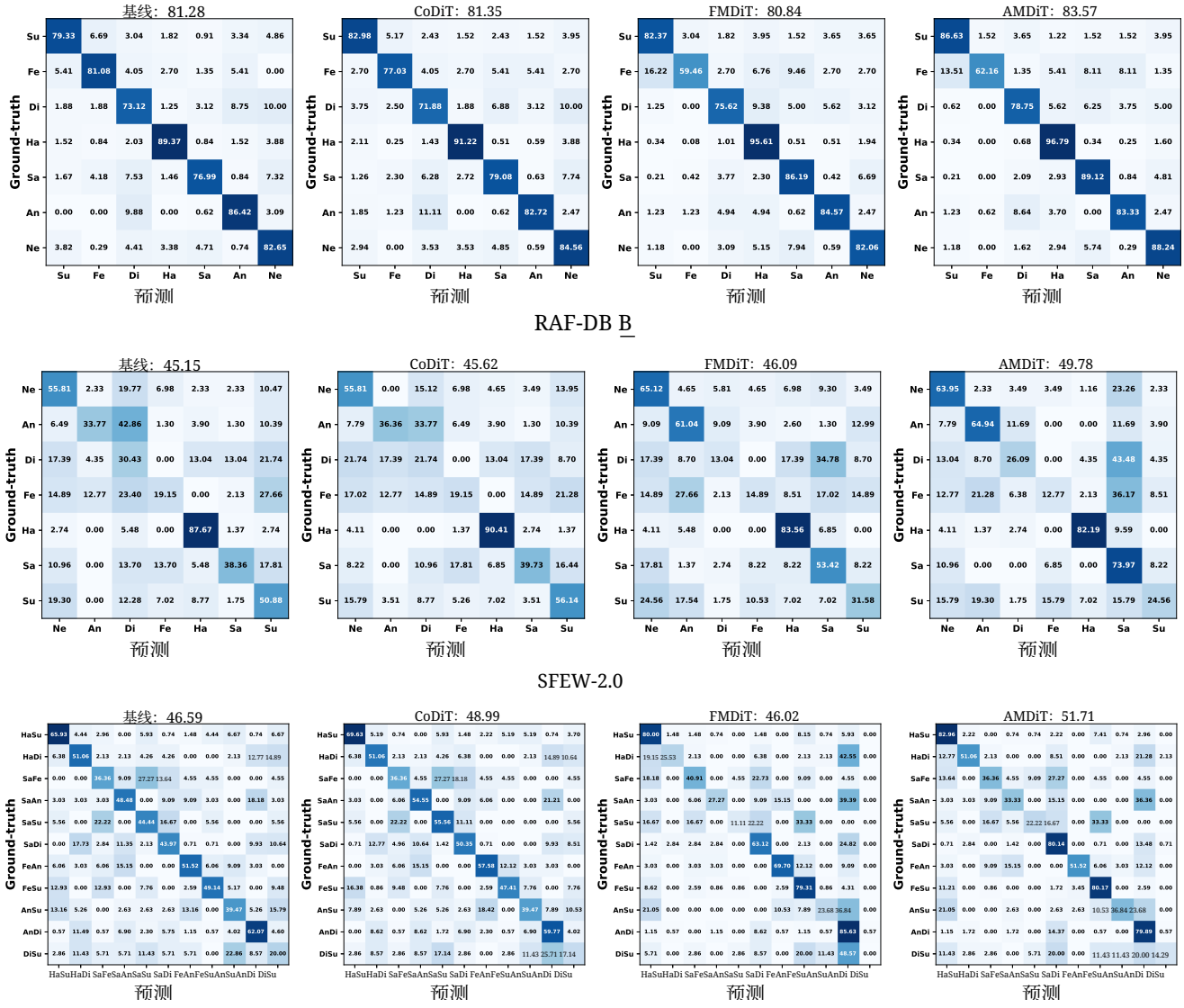


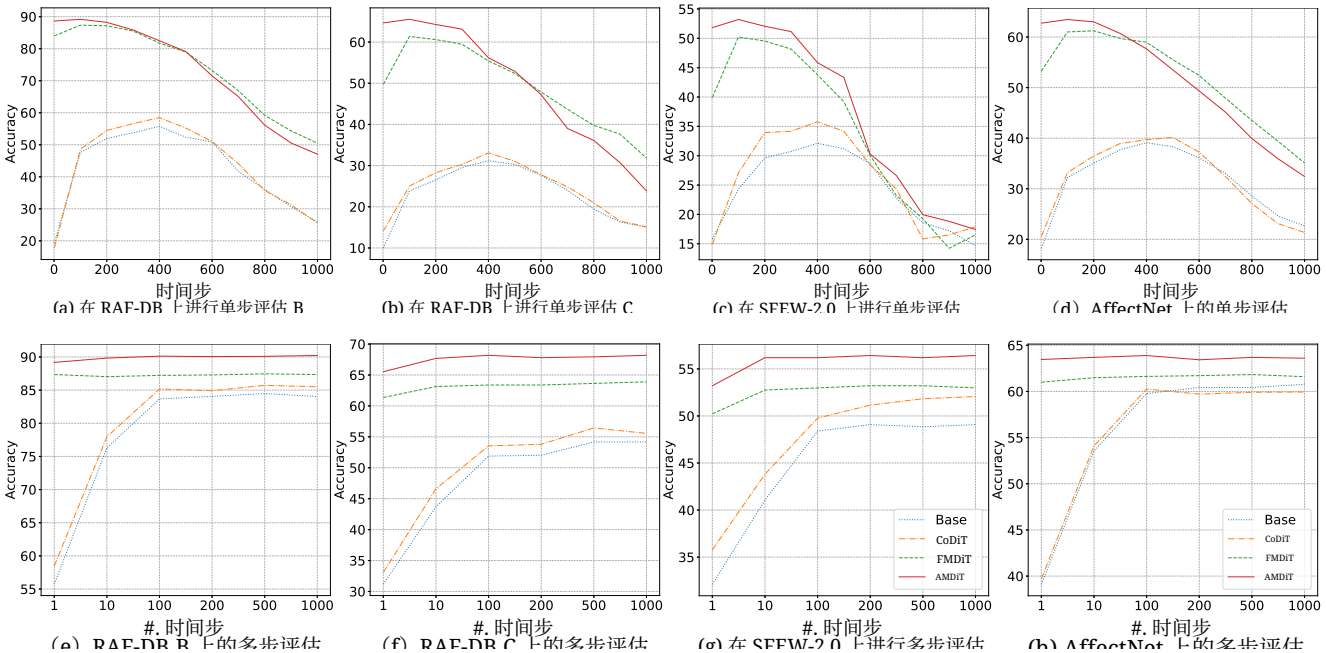
图 5：四种 EmoDC 变体在三个数据集上进行 100 步评估的混淆矩阵。每个矩阵上方显示平均类别准确率（%）。对于 AMDiT，展示的是 $c_{\mathrm{nn}} = c_p$ 的结果。

较于 EmoDC Base，准确率分别达到 89.08%、63.51%、56.42% 和64.49%。值得注意的是，AMDiT($c_{\mathrm{nn}} = c_p$) 在 RAF-DB B 和 RAF-DB C 上的准确率分别提升了 6.45% 和 16.29%，表现优于所有其他方法。

鉴于数据集固有的类别不平衡问题，我们同时评估类别准确率和平均类别准确率，以确保结果适用于实际应用。图 5 使用 100 步评估对比了四种 EmoDC 变体。结果表明，与基线方法相比，CoDiT 提高了平均类别准确率。虽然 FMDiT 提升了整体准确率，但可能会降低平均类别准确率。造成性能下降的因素之一，是固定间隔可能不够大，无法有效处理某些类别中的困难样本，例如 RAF-DB C 中的“快乐-厌恶”类别。相比之下，AMDiT 方法缓解了这一问题，并且显著在平均类别准确率方面以较大优势超越了其他三种 EmoDC 变体。这表明，该方法能够有效提升面部表情识别性能，无论是整体准确率还是平均类别准确率。

4）不同时间步的评估：本节分析单步评估和多步评估。单步评估使用一个时间步获得分类结果，而多步评估则使用多个间隔均匀的时间步（例如 10、100、200）获得结果。评估在全部四个数据集上进行。

单步评估。为阐明所提出方法的有效性和效率，我们分析了 EmoDC 在不同时间步上的单步评估性能。评估涵盖 11 个时间步： $\{1, 100, 200, ..., 900, 1000\}$。图 6(a)-(d) 展示了四种 EmoDC 变体的单步评估结果，由此得出三项关键观察。首先，差异



(a) 在 RAF-DB 上进行单步评估 B (b) 在 RAF-DB 上进行单步评估 C (c) 在 SFEW-2.0 上进行单步评估 (d) AffectNet 上的单步评估

(e) RAF-DB B 上的多步评估 (f) RAF-DB C 上的多步评估 (g) 在 SFEW-2.0 上进行多步评估 (h) AffectNet 上的多步评估

图 6：四种 EmoDC 变体在四个数据集上的单步和多步评估。对于 AMDiT，展示了使用 $c_{\mathrm{nn}} = c_p$ 得到的结果。

性训练相较于基线 EmoDC Base 提升了 EmoDC 在单步评估中的性能。CoDiT 仅带来了小幅准确率提升，而基于间隔的方法则使准确率显著高于 EmoDC Base 和 CoDiT。其次，基于间隔的方法改变了 EmoDC 达到最高准确率的时间步。具体而言，EmoDC Base 和 CoDiT 的最高准确率均出现在约第 400 个时间步，这与文献 [21] 中关于扩散分类器的既有发现一致。在该时间步，EmoDC Base 在 RAF-DB B、RAF-DB C、SFEW-2.0 和 AffectNet 上的准确率分别为 55.74%、31.19%、32.11% 和 39.11%；CoDiT 在相应数据集上的准确率分别为 58.47%、33.08%、35.78% 和 39.71%。相比之下，基于间隔的方法几乎都在更早的时间步（约第 100 个时间步）达到最高准确率。在该时间步，FMDiT 在 RAF-DB B、RAF-DB C、SFEW-2.0 和 AffectNet 上的准确率分别为 87.35%、61.36%、50.23% 和 61.00%；AMDiT 的表现进一步提升，在相同数据集上的准确率分别达到 89.18%、65.53%、53.21% 和 63.46%。第三，FMDiT 和 AMDiT 在单步评估中的峰值准确率显著高于 EmoDC Base。这些发现凸显了所提出的基于间隔的方法在有效性和效率方面的卓越表现，表明它们具备更强的能力来优化 EmoDC 的单步性能。

多步评估。在多个时间步上计算平均噪声预测误差，比仅依赖单个时间步能获得更高的准确率。因此，我们采用不同数量的时间步进行评估，以考察四种 EmoDC 变体的潜力。实验采用单步评估，以及均匀间隔的 10、100、200、500 和 1,000 步评估。对于单步评估，第 400 根据此前的单步评估结果，EmoDC Base 和 CoDiT 使用该时间步，而 FMDiT 和 AMDiT 则采用第 100 个时间步。图 6(e)–(h) 展示了多步评估结果，揭示了两点重要发现。一方面，与 EmoDC Base 相比，差异训练在不同的多步评估中提升了性能。尽管 CoDiT 仅带来了小幅改进，但基于间隔的方法 FMDiT 和 AMDiT 的表现远胜 EmoDC Base 和 CoDiT。这一观测值证实，基于间隔的差异训练在提升 EmoDC 性能方面具有显著优势。另一方面，基于间隔的方法大幅缩小了单步评估与多步评估之间的准确率差距。通常，扩散分类器需要更多时间步才能达到更高的识别准确率 [21]。例如，在 RAF-DB B 数据集上，EmoDC Base 需要 500 个时间步，才能比使用 1 个时间步时的准确率高出 30%，但这需要约 $500\times$的计算成本。基于间隔的方法有效缓解了这一问题。EmoDC Base 和 CoDiT 的准确率在超过 100 个时间步后趋于稳定，且没有显著提升；而 FMDiT 和 AMDiT 使用超过 10 个时间步便能达到稳定状态。此外，FMDiT 和 AMDiT 在单步评估与多步评估之间的性能差距仍然很小。这表明，所提出的基于间隔的方法能够通过减少时间步来加快推理速度，同时不会显著牺牲准确率。值得注意的是，在四个数据集上，基于间隔的方法的单步准确率均超过了 EmoDC Base 和 CoDiT 的多步评估峰值性能。这一性能提升可归因于基于间隔的方法能够在不同时间步上持续提高负向配对中的噪声预测误差，

表IV：最先进的判别分类器和 EmoDC 在 RAF-DB B、SFEW-2.0 和 AffectNet 上的准确率（%）。“FP”表示模型在微调前是否已在人脸图像数据集上进行预训练。

判别式分类器				
方法	FP	RAF-DB B	SFEW-2.0	AffectNet
DLP-CNN [29]	No	84.13	51.05	-
岛损失 [63]	No	-	52.52	-
LBAN-IL [64]	No	85.89	55.28	-
IPA2LT (LTNet) [65]	No	86.77	58.29	57.31
RAN [59]	Yes	86.90	56.40	59.50
GA-FER [3]	Yes	87.52	57.11	-
DACL [66]	Yes	87.78	-	65.20
VTFF [35]	Yes	88.14	-	61.85
EfficientFace [67]	Yes	88.36	-	63.70
Meta-Face2Exp [68]	Yes	88.76	-	64.23
T-VPT-M [6]	No	89.70	58.49	-
APViT [37]	Yes	91.98	**61.92**	66.91
POSTER [69]	Yes	92.05	57.34	67.31
POSTER++ [70]	Yes	**92.21**	57.57	**67.49**
情绪扩散分类器 (EmoDC，我们的方法)				
Base	否	83.67	48.39	59.63
AMDiT ($c_{\text{nn}} = c_{\text{nc}}$)	No	89.08	**56.42**	**64.49**
AMDiT ($c_{\text{nn}} = c_p$)	No	**90.12**	56.19	63.86

表 V：RAF-DB C 上最先进的判别分类器和 EmoDCon 的准确率（%）。“FP”表示模型在微调前是否在人脸图像数据集上进行了预训练。

判别式分类器		
方法	FP	RAF-DB C
DLP-CNN [29]	No	57.95
分离损失 [71]	No	58.84
ESG-Net [72]	Yes	61.90
CDNet [73]	Yes	63.03
双中心损失 [4]	Yes	66.16
T-VPT-M [6]	No	67.17
情绪扩散分类器（EmoDC，我们的方法）		
基线	无	51.89
AMDiT ($c_{\text{nn}} = c_{\text{nc}}$)	No	63.51
AMDiT ($c_{\text{nn}} = c_p$)	No	**68.18**

从而增强 EmoDC 区分不同面部表情的能力。AMDiT 在所有数据集和多步评估中始终优于 FMDiT，凸显了 EmoDC 训练中采用的自适应间隔方法的有效性和效率。

5）与判别式分类器的比较：本节评估采用所提出的 AMDiT 方法训练的 EmoDC 在四个真实场景面部表情数据集上的表现。作为比较，我们将 EmoDC 与最先进（SOTA）的判别式分类器进行基准测试。表 IV 和表 V 展示了结果，并指出模型在重新训练前是否曾使用面部数据集进行预训练。使用面部图像进行预训练通常能够增强模型进行面部属性分析的能力，并提升其面部表情识别性能。在三个面部表情数据集上（不包括 AffectNet），EmoDC Base 方法的表现均不及任何 SOTA 方法。不过，采用所提出的 AMDiT 后，EmoDC 在 RAF-DB B 和 AffectNet 上的表现超过了若干先前的 SOTA 方法，例如 LBAN-IL [35], RAN [59], VTFF [35],和 Meta-Face2Exp [68]。尽管如此，EmoDC 在 RAF-DB B 上仍不及最强的 SOTA 模型，例如 APViT [37], POSTER [69], 和 POSTER++ [70]，不过准确率差距相对较小。对于复合面部表情数据集 RAF-DB C，表 V 显示，采用 AMDiT 的 EmoDC 达到了 68.18% 的 SOTA 准确率。尽管所提出的采用 AMDiT 的 EmoDC 尚未超越基础面部表情数据集上顶尖的 SOTA 判别式分类器，但它通过利用扩散模型的生成能力，开创了一种新的面部表情识别范式。尽管该方法尚未取得基准性能优势，却为拓展面部表情识别研究及其实际应用范围奠定了创新基础。

6）鲁棒性评估：判别式分类器容易受到捷径学习和对抗样本的影响，而生成式扩散分类器则表现出更强的鲁棒性[23], [24]。本节评估 EmoDC 在噪声和模糊条件下的鲁棒性。EmoDC 在干净图像上进行微调，并在对抗样本上进行评估。具体而言，在不同噪声水平下，向原始测试图像添加高斯噪声 $\epsilon_{\text{img}} \sim (\mu,\ \sigma_n^2)$，其中 $\mu = 0$ 和 $\sigma_n = \{10, 20, 30, 40, 50\}$。值得注意的是，虽然高斯噪声是直接添加到原始图像中的，但 Stable Diffusion 会在潜在空间中自然地引入并移除噪声 [15]。通过使用大小为 15、标准差为 $\sigma_b = \{2, 5, 10, 15\}$ 的高斯核合成模糊图像。为评估性能，我们将 EmoDC 与三个开源 SOTA 模型进行比较：APViT [37],POSTER [69], 和 POSTER++[70]。

表 VI 和表 VII 所示的实验结果表明，与 POSTER 和 POSTER++ 相比，APViT 对噪声样本表现出更强的鲁棒性；而与 APViT 相比，POSTER 和 POSTER++ 在处理模糊的面部表情图像方面表现更佳。尽管 POSTER 和 POSTER++ 在干净图像上取得了更高的准确率，但随着噪声水平升高，其性能显著下降。这种下降可能是因为在噪声条件下，关键点检测的有效性降低。相比之下，APViT 对噪声的耐受性更强，但在模糊程度较高时（σ_b）性能会急剧下降。反过来，POSTER 和 POSTER++ 对模糊的耐受性优于 APViT。所提出的 EmoDC 变体在噪声和模糊条件下均表现出优异性能。EmoDC Base 在干净图像上的表现起初不及 SOTA 方法，但随着噪声或模糊程度增加，仍能保持相对稳定的性能。融入差异训练的增强变体进一步提升了基线模型的性能。当噪声水平超过 $\sigma_n = 20$ 或模糊程度超过 $\sigma_b = 5$ 时，AMDiT 变体对噪声和模糊均表现出很强的耐受性，优于所有其他方法。虽然三种表现优异的 SOTA 判别式分类器对噪声或模糊的鲁棒性有限，但采用 AMDiT 的 EmoDC 在这两种条件下始终表现出色且稳健。这些发现凸显了 EmoDC 借助 AMDiT 有效缓解噪声和模糊条件下性能下降的能力，使其成为一种稳健且实用的解决方案，可用于真实场景中的面部表情识别，且无需额外的图像恢复模型。

7）跨数据集评估：我们通过在 RAF-DB B 上训练模型，并在以下数据集上进行测试来开展跨数据集评估

表 VI：评估 EmoDC 在 RAF-DB B 和 SFEW-2.0 上对噪声的鲁棒性，其中噪声为高斯噪声（$\mu = 0$, $\sigma_n = \{10, 20, 30, 40, 50\}$）。“无噪声”表示不含噪声的图像。

方法	RAF-DB B						SFEW-2.0*					
	无噪声	10	20	30	40	50	无噪声	10	20	30	40	50
APViT [37]	91.98	**90.74**	87.71	84.55	78.94	70.70	57.34	54.82	45.41	42.43	33.49	26.15
POSTER [69]	92.05	90.55	87.16	79.53	66.23	46.22	57.34	50.00	45.18	35.32	27.98	25.23
POSTER++ [70]	**92.21**	89.67	85.66	73.24	54.79	35.72	**57.57**	46.56	36.24	25.00	20.18	19.73
Base	83.67	79.47	79.04	76.86	73.99	70.93	48.39	49.08	48.62	46.56	45.18	**43.35**
CoDiT	85.17	80.08	80.15	77.71	75.39	72.16	49.77	49.31	46.10	44.27	43.35	37.84
FMDiT	87.22	86.08	85.37	83.47	82.66	80.41	52.98	50.69	46.33	40.83	37.16	33.26
AMDiT (c_{nc})	89.08	88.62	87.94	86.51	83.31	80.44	56.42	**55.05**	**52.98**	**49.54**	**46.33**	41.28
AMDiT (c_p)	90.12	89.86	**88.46**	**86.99**	**85.37**	**84.19**	56.12	54.82	49.77	46.56	44.27	41.28

*APViT、POSTER 和 POSTER++ 模型均使用与我们的方法相同的对齐图像重新训练。

表 VII：使用核大小为 15、标准差为 $\sigma_b = \{2, 5, 10, 15\}$ 的高斯模糊，评估 EmoDC 在 RAF-DB B 和 SFEW-2.0 上对模糊的鲁棒性。

方法	RAF-DB B					SFEW-2.0*				
	无噪声	2	5	10	15	无噪声	2	5	10	15
APViT [37]	91.98	78.94	42.37	38.79	38.56	57.34	52.52	25.92	24.08	23.39
POSTER [69]	92.05	**90.65**	**82.17**	78.68	78.10	57.34	**56.65**	49.54	45.64	44.72
POSTER++ [70]	**92.21**	90.35	81.06	78.52	78.06	**57.57**	55.05	47.48	43.12	41.92
Base	83.67	61.80	57.46	60.72	61.34	48.39	47.02	48.85	48.85	47.71
CoDiT	85.17	73.21	66.53	66.40	66.85	49.77	48.85	49.77	48.62	47.94
FMDiT	87.22	83.15	76.37	74.77	74.87	52.98	53.67	50.46	47.48	47.25
AMDiT (c_{nc})	89.08	88.85	82.14	**80.22**	**79.76**	56.42	56.42	**56.88**	**56.42**	**55.96**
AMDiT (c_p)	90.12	87.74	78.68	76.92	76.50	56.12	53.67	50.92	47.94	47.94

*APViT、POSTER，a 以及 POSTER++ 模型均使用与我们的方法相同的对齐 图像重新训练。



图 7：使用在 RAF-DB B 上训练的 EmoDC 与 AMDiT（$c_{\text{nn}} = c_p$）生成面部表情的示例。图像是在无分类器引导尺度（cfg）为 3 和 5 的条件下生成的。

表 VIII：在 SFEW-2.0、FER2013Plus 和 RAF-DB C 上的跨数据集评估结果。所有模型均在 RAF-DB B 训练集上训练。

方法	SFEW-2.0	FER2013Plus	RAF-DB C
APViT [37]	51.83	72.71	-
POSTER [69]	44.95	63.52	-
POSTER++ [70]	48.85	66.50	-
Base	48.17	64.99	43.94
AMDiT (c_{nc})	51.72	77.13	48.48
AMDiT (c_p)	**52.06**	**78.35**	**53.91**

：SFEW-2.0、FER2013Plus 和 RAF- DB C。值得注意的是，三种先进的判别式分类器表现出有限的跨数据集泛化能力，尤其是在从基本面部表情转向复合面部表情时。相比之下，EmoDC 在此类评估中展现出稳健的性能。这一优势源于判别式分类器受限于预定义的类别，而基于扩散的分类器能够更有效地泛化到未见类别。表 VIII 所示结果表明，AMDiT 提升了 EmoDC 的跨数据集性能，优于 EmoDC 基础模型以及三种强大的判别式分类器。这些发现凸显了 AMDiT 在提升跨数据集泛化能力方面的有效性。

8）生成分析：生成式扩散分类器的一项关键优势在于能够生成视觉图像，从而更清晰地了解模型所习得的知识。与判别式分类器相比，这一能力提升了模型的可解释性。扩散模型通过迭代去噪过程生成图像，过程始于从高斯分布中采样的随机噪声 [14], [15], [55]。本研究使用带有 AMDiT 的 EmoDC 变体生成面部表情图像（$c_{\text{nn}} = c_p$）。如图 7 所示，这些图像展示了在两个无分类器引导尺度（cfg）3 和 5 下合成的七种基本面部表情。这种生成能力凸显了模型对面部表情底层数据分布的深刻理解。cfg 尺度控制生成多样性与保真度之间的平衡 [74]：较高的 cfg 值会提高保真度，但会牺牲多样性，反之亦然。在本例中，较大的 cfg 值会生成更明显的面部表情，其特征是面部肌肉运动更加夸张。具体而言，使用 cfg = 5 生成的图像呈现出更鲜明的形状——这是识别面部表情中情绪状态的关键指标。这一发现与 [49], 一致，后者指出扩散分类器具有类似人类的形状偏差。面部表情之所以能够被精确生成，是因为生成式扩散模型准确地建模了真实数据分布，这也有助于 EmoDC 获得较高的准确率。使用定义清晰且多样的面部表情进行训练，能增强模型辨别细微类间差异（例如区分“惊讶”和“恐惧”）以及类内变化（例如“悲伤”的程度）的能力。然而，使用 AMDiT 训练的 EmoDC 在生成的面部表情方面暴露出两项

表IX：判别式模型和 EmoDC 变体的 GPU 训练时长与推理速度（包括单图推理时间和每秒帧数（FPS））比较。所有评估均在 RTX 4090 GPU 上进行。

方法	GPU 小时数	推理时间（毫秒/张图像）	FPS
APViT [37]	0.8	7.8	128
POSTER [69]	4.6	14.7	68
POSTER++ [70]	2.5	5.2	192
EmoDC（1 步）	8.6	29.5	34
EmoDC（10 步）	8.6	143.0	7

个显著局限：色调不自然，以及会降低视觉连贯性的异常模式。尽管存在这些局限，本研究的重点仍是面部表情识别。如何同时提升生成和识别性能，是未来研究的一个方向。

V. 结论与讨论

A. 结论

本文探讨了 Stable Diffusion 在面部表情识别（FER）中的应用，并提出了情绪扩散分类器（EmoDC）。为提升其性能，我们首先采用差异训练，促使模型在正类描述条件下准确预测噪声，同时放大其在负类描述条件下的噪声预测误差。在此基础上，我们提出基于间隔的差异训练，在正、负图文对的噪声预测误差之间强制设定最小间隔。然而，固定间隔无法适应不同样本的噪声预测难度。为解决这一局限，我们提出自适应间隔差异训练（AMDiT），通过动态间隔为每个样本量身定制训练方案。实验结果表明，AMDiT 显著提升了其在四个面部表情识别数据集——RAF-DB B、RAF-DB C、SFEW-2.0 和 AffectNet——上的准确率。AMDiT 不仅实现了更高的准确率，还缩小了单步评估与多步评估之间的准确率差距，从而加快推理过程。此外，AMDiT 提升了 EmoDC 在噪声和模糊条件下的鲁棒性，并改善了跨数据集性能，优于当前最先进的判别式分类器。

B. 讨论

尽管 AMDiT 使 EmoDC 相较于最先进的判别模型具备更强的鲁棒性和跨数据库泛化能力，但这也带来了更高的计算开销。如表IX所示，训练和推理时间均长于判别模型，主要原因有两点：（i）模型规模较大；（ii）需要针对不同类别和时间步进行多次模型前向计算。具体而言，EmoDC 利用了功能强大的预训练 Stable Diffusion 模型，该模型最初是为高分辨率图像生成而设计的 [15]。在推理过程中，每张图像都会由 VAE 编码器编码以生成潜在表示，随后由 UNet 多次处理，以评估不同的类别提示和时间步。这些过程共同导致推理速度变慢。尽管存在这些限制，本研究旨在验证强大生成模型应用于面部表情识别的可行性与潜力，并提出了新颖的 AMDiT 方法，以进一步提升 EmoDC 的性能和鲁棒性。我们承认，优化网络规模和推理速度仍是未来工作的关键方向。重要的是，EmoDC 已具备实际部署的可行性。例如，在单步评估设置下，EmoDC 的吞吐量达到 34 帧每秒，与 EmoDCBase 相比，准确率没有显著下降，同时能够实现实时面部表情识别。与此同时，10步评估提供了准确率更高的替代方案，适用于低帧率摄像头或对速度要求不高的高精度系统。此外，尽管本研究聚焦于静态面部表情识别，将生成式分类器扩展到动态的基于视频的识别，是建模情绪时间变化的重要方向。

参考文献

[1] S. Li 和 W. Deng，《深度面部表情识别：综述》，*IEEE*情感计算汇刊，第13卷第3期，页1195–1215，2022年。[2] Y. Wang、W. Song、W. Tao、A. Liotta、D. Yang、X. Li、S. Gao、Y. Sun、W. Ge、W. Zhang 等，《情感计算系统综述：情绪模型、数据库与最新进展》，信息融合，第83卷，页19–52，2022年。[3] R. Zhao、T. Liu、Z. Huang、D. P. Lun 和 K.-M. Lam，《通过注意力图卷积网络实现几何感知面部表情识别》，*IEEE*情感计算汇刊，第14卷，页1174，2023年2，1159第期，页–。[4] R. Dong 和 K.-M. Lam，《用于复合面部表情识别的双中心损失》，*IEEE*信号处理快报，页641–645，2024年。[5] ——，《利用基本表情数据进行复合面部表情识别的半监督学习》，载于国际先进成像技术研讨会，第13164卷，2024年，页314–319。[6] R. Dong、C. Yang 和 K.-M. Lam，《利用掩码图像进行文本引导的视觉提示调优，以实现面部表情识别》，载于亚太信号与信息处理协会年度峰会及会议论文集，2024年，页1–6。[7] K. He、X. Zhang、S. Ren 和 J. Sun，《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，2016年，页770–778。[8] C. Guo、G. Pleiss、Y. Sun 和 K. Q. Weinberger，《现代神经网络的校准问题》，载于国际机器学习会议论文集，2017年，页1321–1330。[9] R. Geirhos、J.-H. Jacobsen、C. Michaelis、R. Zemel、W. Brendel、M. Bethge 和 F. A. Wichmann，《深度神经网络中的捷径学习》，自然机器智能，第2卷第11期，页665–673，2020年。[10] J. Robinson、L. Sun、K. Yu、K. Batmanghelich、S. Jegelka 和 S. Sra，《对比学习能否避免捷径解法？》，载于神经信息处理系统进展会议论文集，第34卷，2021年，页4974–4986。[11] A. C. Li、A. Kumar 和 D. Pathak，《生成式分类器可避免捷径解法》，载于国际学习表征会议论文集，2025年。[12] A. Nguyen、J. Yosinski 和 J. Clune，《深度神经网络很容易被欺骗：对不可识别图像做出高置信度预测》，载于*IEEE*计算机视觉与模式识别会议论文集，2015年，页427–436。[13] A. Ng 和 M. Jordan，《判别式分类器与生成式分类器之比较：逻辑回归与朴素贝叶斯》，载于神经信息处理系统进展会议论文集，第14卷，2001年。[14] J. Ho、A. Jain 和 P. Abbeel，《去噪扩散概率模型》，载于神经信息处理系统进展会议论文集，第33卷，2020年，页6840–6851。[15] R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer，《利用潜在扩散模型进行高分辨率图像合成》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，页10 684–10 695。[16] W. Peebles 和 S. Xie，《利用Transformer实现可扩展扩散模型》，载于*IEEE/CVF*国际计算机视觉会议论文集，2023年，页4195–4205。[17] L. Qu、W. Wang、Y. Li、H. Zhang、L. Nie 和 T.-S. Chua，《用于文生图的判别式探测与调优》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，页7434–7444。

[18] C. Yang、R. Dong 和 K.-M. Lam，《真实场景全景图像超分辨率的高效适配》，载于亚太信号与信息处理协会年度峰会及会议，2024年，页1–6。[19] Y. Zuo、J. Xiao、K.-C. Chan、R. Dong、C. Yang、Z. He、H. Xie 和 K.-M. Lam，《通过视觉条件实现单步扩散的多视图一致性风格迁移》，载于欧洲计算机视觉会议研讨会论文集，2024年。[20] C. Yang、R. Dong 和 K.-M. Lam，《视觉语言模型引导的图像复原》，arXiv预印本 arXiv:2512.17292，2025年。[21] A. C. Li、M. Prabhudesai、S. Duggal、E. Brown 和 D. Pathak，《你的扩散模型其实是一个零样本分类器》，载于IEEE/CVF国际计算机视觉会议论文集，2023年，页2206–2217。[22] K. Clark 和 P. Jaini，《文生图扩散模型是零样本分类器》，载于神经信息处理系统进展会议论文集，第36卷，2023年，页58 921–58 937。[23] H. Chen、Y. Dong、Z. Wang、X. Yang、C. Duan、H. Su 和 J. Zhu，《基于单一扩散模型的鲁棒分类》，载于国际机器学习会议论文集，第257期，2024年，页6643–6665。[24] H. Chen、Y. Dong、S. Shao、H. Zhongkai、X. Yang、H. Su 和 J. Zhu，《扩散模型是可认证鲁棒分类器》，载于神经信息处理系统进展会议论文集，第37卷，2024年，页50 062–50 097。[25] E. J. Hu、Y. Shen、P. Wallis、Z. Allen-Zhu、Y. Li、S. Wang、L. Wang 和 W. Chen，《LoRA：大语言模型的低秩适配》，载于国际学习表征会议论文集，2022年。[26] K. Simonyan 和 A. Zisserman，《用于大规模图像识别的超深卷积网络》，载于国际学习表征会议论文集，2015年，页1–10。[27] J. H. Kim、A. Poulose 和 D. S. Han，《CVGG-19：用于面部情绪识别的定制视觉几何组深度学习架构》，IEEE Access，第12卷，页41 557–41 578，2024年。[28] ——，《面部图像脱粒机在面部情绪识别性能中的广泛应用》，Sensors，第21卷，2026年，2021年6，页。[29] S. Li、W. Deng 和 J. Du，《可靠的众包与深度局部保持学习：用于野外表情识别》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2017年，页2852–2861。[30] Y. Li、J. Zeng、S. Shan 和 X. Chen，《使用带注意力机制的卷积神经网络进行遮挡感知面部表情识别》，IEEE图像处理汇刊，第28卷第5期，页2439–2450，2018年。[31] R. Zhao、T. Liu、Z. Huang、D. P. Lun 和 K.-M. Lam，《时空图与Transformer结合的几何引导面部表情识别》，IEEE情感计算汇刊，第14卷，页2767，2022年4，2751第期，页–。[32] W. Chen、D. Zhang、M. Li 和 D.-J. Lee，《Stcam：用于动态面部表情识别的时空与通道注意力模块》，IEEE情感计算汇刊，第14卷第1期，页800–810，2020年。[33] C. Gan、J. Xiao、Z. Wang、Z. Zhang 和 Q. Zhu，《使用密集连接卷积神经网络和层次空间注意力进行面部表情识别》，图像与视觉计算，第117卷，页104342，2022年。[34] F. Xue、Q. Wang 和 G. Guo，《TransFER：利用Transformer学习关系感知面部表情表征》，载于IEEE/CVF国际计算机视觉会议论文集，2021年，页3601–3610。[35] F. Ma、B. Sun 和 S. Li，《利用视觉Transformer和注意力选择性融合进行面部表情识别》，IEEE情感计算汇刊，第14卷第2期，页1236–1248，2021年。[36] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly、J. Uszkoreit 和 N. Houlsby，《一张图像胜过16×16个词：用于大规模图像识别的Transformer》，载于国际学习表征会议论文集，2021年。[37] F. Xue、Q. Wang、Z. Tan、Z. Ma 和 G. Guo，《利用注意力池化的视觉Transformer实现鲁棒面部表情识别》，IEEE情感计算汇刊，第14卷第4期，页3244–3256，2022年。[38] Z. Zhao 和 I. Patras，《提示视觉语言模型进行动态面部表情识别》，载于英国机器视觉会议论文集，2023年。[39] H. Zhou、S. Huang、F. Zhang 和 C. Xu，《Ceprompt：用于面部表情识别的跨模态情绪感知提示》，IEEE电路与系统视频技术汇刊，2024年。[40] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark等，《通过自然语言监督学习可迁移的视觉模型》，载于国际机器学习会议论文集，2021年，页8748–8763。[41] Y. L. Bouali、O. B. Ahmed 和 S. Mazouzi，《表演式语音的音视频情绪识别跨模态学习》，载于信号与图像处理先进技术国际会议论文集，2022年，页1–6。

[42] E. Ryumina, M. Markitantov, D. Ryumin, H. Kaya, and A. Karpov, “基于情绪概率融合的零样本视听复合表情识别方法”，载于 IEEE/CVF 计算机视觉与模式识别会议研讨会论文集，2024年，第 4752–4760 页。[43] S. Du, Y. Tao, and A. M. Martinez，“情绪的复合面部表情”，《美国国家科学院院刊》，第 111 卷，第 15 期，第 E1454–E1462 页，2014年。[44] E. Fetaya, J.-H. Jacobsen, W. Grathwohl, and R. Zemel，“理解条件生成模型的局限性”，载于国际学习表征会议论文集，2020年。[45] R. Mackowiak, L. Ardizzone, U. Kothe, and C. Rother，“以生成式分类器为基础实现可信图像分类”，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2021年，第 2971–2981 页。[46] S. Mukhopadhyay, M. Gwilliam, V. Agarwal, N. Padmanabhan, A. Swaminathan, S. Hegde, T. Zhou, and A. Shrivastava，“扩散模型在图像分类方面优于生成对抗网络”，arXiv 预印本 arXiv:2307.08702，2023 年。[47] P. Dhariwal and A. Nichol，“ 扩散模型在图像合成方面优于生成对抗网络”，载于神经信息处理系统进展论文集，第 34 卷，2021年，第 8780–8794 页。[48] X. Han, H. Zheng, and M. Zhou，“CARD：分类与回归扩散模型”，载于神经信息处理系统进展论文集，第 35 卷，2022年，第 18 100–18 115 页。[49] P. Jaini, K. Clark, and R. Geirhos，“生成式分类器的有趣特性”，载于国际学习表征会议论文集，2024年。[50] K. Q. Weinberger and L. K. Saul，“用于大间隔最近邻分类的距离度量学习”，《机器学习研究期刊》，第 10 卷，第 2 期，2009年。[51] F. Schroff, D. Kalenichenko, and J. Philbin，“FaceNet：用于人脸识别和聚类的统一嵌入”，载于 IEEE 计算机视觉与模式识别会议论文集，2015年，第 815–823 页。[52] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl，“采样对深度嵌入学习至关重要”，载于 IEEE 国际计算机视觉会议论文集，2017年，第 2840–2848 页。[53] O. Rippel, M. Paluri, P. Dollar, and L. Bourdev，“采用自适应密度判别的度量学习”，载于国际学习表征会议论文集，2015年。[54] K. Sohn，“采用多类 N 对损失目标改进深度度量学习”，载于神经信息处理系统进展论文集，第 29 卷，2016年。[55] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli，“利用非平衡热力学进行深度无监督学习”，载于国际机器学习会议论文集，2015年，第 2256–2265 页。[56] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon，“复杂条件下的静态面部表情分析：数据、评估协议与基准”，载于 IEEE 国际计算机视觉会议研讨会，2011年，第 2106–2112 页。[57] A. Mollahosseini, B. Hasani, and M. H. Mahoor，“AffectNet：用于自然场景下面部表情、效价和唤醒度计算的数据库”，《IEEE 情感计算汇刊》，第 10 卷，第 1 期，第 18–31 页，2017年。[58] E. Barsoum, C. Zhang, C. Canton Ferrer, and Z. Zhang，“利用众包标签分布训练面部表情识别深度网络”，载于 ACM 多模态交互国际会议论文集，2016年。[59] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao，“用于应对姿态变化和遮挡的区域注意力网络面部表情识别”，《IEEE 图像处理汇刊》，第 29 卷，第 4057–4069 页，2020年。[60] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee 等，“表征学习的挑战：三项机器学习竞赛报告”，载于神经信息处理国际会议论文集，Springer，2013年，第 117–124 页。[61] I. Loshchilov and F. Hutter，“解耦权重衰减正则化”，载于国际学习表征会议论文集，2019年。[62] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao，“使用多任务级联卷积网络进行联合人脸检测与对齐”，《IEEE 信号处理快报》，第 23 卷，第 10 期，第 1499–1503 页，2016年。[63] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O’Reilly, and Y. Tong，“用于面部表情识别判别特征学习的岛损失”，载于 IEEE 自动化人脸与手势识别国际会议论文集，IEEE，2018年，第 302–309 页。[64] H. Li, N. Wang, Y. Yu, X. Yang, and X. Gao，“ LBAN-IL：一种用于面部表情识别的高判别性表征新方法”，《神经计算》，第 432 卷，第 159–169 页，2021年。[65] J. Zeng, S. Shan, and X. Chen，“利用标注不一致的数据集进行面部表情识别”，载于欧洲计算机视觉会议论文集，2018年，第 222–237 页。[66] A. H. Farzaneh and X. Qi，“通过深度注意力中心损失实现自然场景下面部表情识别”，载于 IEEE/CVF 冬季计算机视觉应用会议论文集，2021年，第 2402–2411 页。

[67] Z. Zhao, Q. Liu, and F. Zhou，“采用标签分布训练的鲁棒轻量级面部表情识别网络”，载于 *AAAI* 人工智能会议论文集，第 35 卷，第 4 期，2021年，第 3510–3519 页。[68] D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang，“Face2exp：应对面部表情识别中的数据偏差”，载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集，2022年，第 20 291–20 300 页。[69] C. Zheng, M. Mendieta, and C. Chen，“POSTER：一种用于面部表情识别的金字塔交叉融合 Transformer 网络”，载于 *IEEE/CVF* 国际计算机视觉会议论文集，2023年，第 3146–3155 页。[70] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, A. Huang, and Y. Wang，“POSTER++：一种更简单、更强大的面部表情识别网络”，模式识别，第 157 卷，文章编号 110951，2025年。[71] Y. Li, Y. Lu, J. Li, and G. Lu，“用于自然场景下基本及复合面部表情识别的分离损失”，载于 亚洲机器学习会议论文集，2019年，第 897–911 页。[72] X. Zou, Y. Yan, J.-H. Xue, S. Chen, and H. Wang，“当面部表情识别遇上少样本学习：一种联合交替学习框架”，载于 *AAAI* 人工智能会议论文集，第 36 卷，第 5 期，2022年，第 5367–5375 页。[73] ——，“Learn-to-decompose：用于跨域少样本面部表情识别的级联分解网络”，载于 欧洲计算机视觉会议论文集，2022年，第 683–700 页。[74] J. Ho and T. Salimans，“无分类器扩散引导”，载于 神经信息处理系统进展研讨会论文集，2021年。