提升生成对抗网络在面部表情合成中的泛化能力

Arbish Akram*、Nazar Khan*、Arif Mahmood†
*旁遮普大学计算机科学系，巴基斯坦拉合尔
†信息技术大学计算机科学系，巴基斯坦拉合尔

arXiv:2603.15648v1 [cs.CV] 4 Mar 2026

摘要—面部表情合成旨在生成逼真的面部表情，同时保持身份一致性。现有的条件生成对抗网络（GAN）在图像到图像转换方面表现出色，但当测试图像与训练数据集存在差异时，其性能往往会下降。我们提出回归生成对抗网络（RegGAN），这是一种通过学习中间表示来提升模型在训练分布之外泛化能力的方法。RegGAN由两个组件构成：一是具有局部感受野的回归层，通过最小化重建误差并采用岭回归损失来学习表情细节；二是通过对抗训练来增强生成图像真实感的细化网络。我们在CFEE 数据集上训练RegGAN，并在CFEE 数据集以及具有挑战性的分布外图像上评估其泛化性能，这些图像包括名人照片、肖像、雕像和虚拟形象渲染图。评估时，我们采用四种广泛使用的指标：用于衡量表情质量的表情分类分数（ECS）、用于衡量身份保持的面部相似度分数（FSS）、用于衡量感知真实感的QualiCLIP，以及用于评估表情质量和真实感的Fr´echet Inception Distance（FID）。在ECS、FID和QualiCLIP方面，RegGAN优于六种最先进模型，在FSS方面则排名第二。人工评估表明，在表情质量、身份保持和真实感方面，RegGAN分别比表现最佳的竞争模型高出25%、26%和30%。

索引词—面部表情合成、生成对抗网络、分布外、细化网络、岭回归层。

I. 引言

面部表情合成旨在将给定表情转换为目标表情。近年来，这一领域取得了显著进展，尤其是生成对抗网络（GAN）[1], [2] 以及近期出现的扩散模型 [3], [4]‚[5]。GAN 支持通过高效的对抗训练实现可控图像生成，而扩散模型则依靠迭代去噪来提升保真度和多样性。在 GAN 中，

变体，StyleGAN [6], [7]已被广泛用于面部表情操控。首先将输入图像投影到StyleGAN潜空间中，在其中执行语义编辑。随后，将修改后的潜代码映射回图像空间。尽管取得了这些进展，FES面临的关键挑战是同时满足以下三个要求：i）表情正确性——合成图像应准确反映目标表情；ii）身份保持——主体身份应保持不变；iii）面部细节保留——输入图像的细微面部细节应得到保留。然而，现有方法难以同时满足所有这些要求。

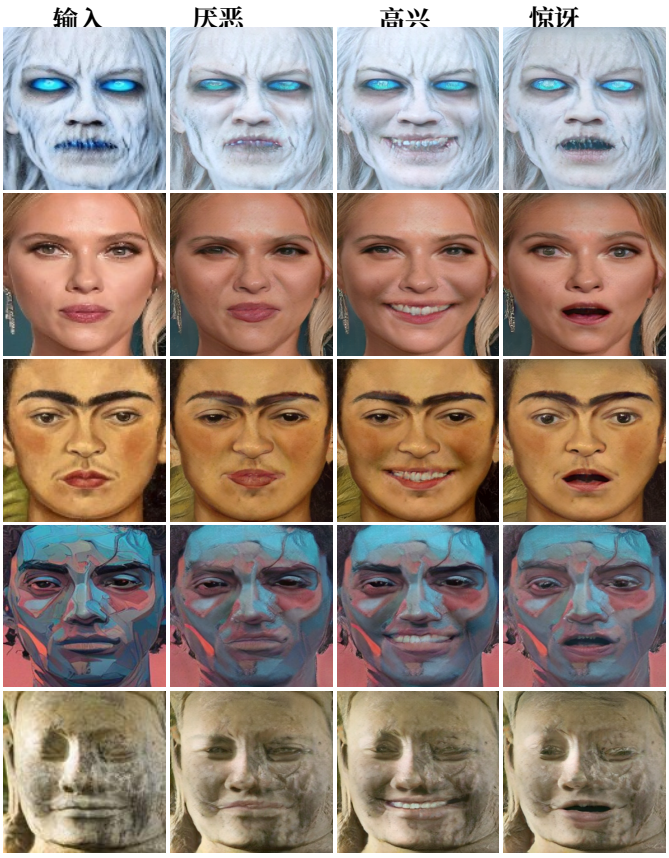

图1：给定一张中性输入图像，所提出的Reg-GAN能够在分布外测试图像上合成照片级真实感的面部表情。尽管仅使用来自CFEE数据集 [8], 的人脸进行训练，所提出的方法仍能引入逼真表情、保持身份并保留输入图像的面部细节。

在这项工作中，我们提出了回归生成对抗网络（RegGAN），这是一种结合回归与对抗式细化的表情合成框架。通过最小二乘目标训练的回归层可生成粗略的表情预测，随后由生成对抗网络对其进行细化，从而提升真实感并保留面部细节。

普通回归层的参数数量会迅速增加，因此难以适应更大的图像尺寸。为了解决这一问题，我们采用了作用于局部图像块的稀疏化岭回归层，大幅减少了

参数数量，同时生成表示目标表情的中间表示。细化网络采用编码器-解码器结构，将该中间表示依次输入多尺度的三类空间注意力块，以增强真实感并保留面部细节。与普通残差块相比，采用空间注意力能让网络更好地捕捉表情和细微的面部特征。RegGAN 分两个阶段依次训练。首先，回归层生成编码目标表情的中间表示。随后，将生成的图像输入细化网络，并通过对抗损失进行训练，以进一步增强真实感。这样的顺序训练使 RegGAN 在表情准确度、身份保持和面部细节方面的表现优于现有方法。我们在分布外人脸图像上评估 RegGAN，包括肖像、虚拟形象、雕像和动物，结果表明，它能够在任何类似面孔的输入上生成类人表情。

本工作的主要贡献如下：

- 将局部化回归变体与对抗学习相结合。这种针对性的回归条件显著提升了生成对抗网络的泛化能力，尤其是在肖像、虚拟形象、奇幻艺术作品、插画、厚涂画和雕像等具有挑战性的图像类型上。
- 采用包含多尺度注意力残差块的细化网络，增强全局和局部面部特征，从而提升生成表情的真实感和保真度。
- 对回归网络和细化网络采用序贯训练策略，使其能够稳健地处理分布外类人脸输入。

本文结构安排如下。第二节简要概述生成对抗网络和最先进的面部表情合成模型。第三节介绍所提出方法的细节。第四节介绍所提方法的数据集与实现细节，以及定性和定量结果与消融研究。第五节分析所提方法的优势和不足，第六节讨论其失败案例。第七节总结全文。

II. 相关工作

近年来，随着生成对抗网络的出现，图像到图像转换任务已能取得照片般逼真的结果 [9], [10], [11], [12], [13], [14], [15], [16], [17]。Pix2pix [9] 引入了一种条件对抗网络，利用 ℓ_1 和对抗损失来学习配对域之间的映射。为解决配对数据对齐问题，CycleGAN [10]通过循环一致性实现了无监督图像到图像转换。这两种方法都已应用于面部表情合成 [18],，但在分布外图像上的性能会下降。

StarGAN [11]通过单一的生成器-判别器架构实现多域面部表情合成。GANimation [19]能够生成不同强度的面部表情，用于逼真动画，但在推理时需要将动作单元作为输入。Cascade-EFGAN [20] 结合局部和全局注意力机制以及渐进式编辑策略，以更好地保留面部细节并减少伪影。研究者还探索了域自适应转换，以生成不同图像风格下的表情，例如素描、油画和动物面孔 [21]。最近，USGAN [22] 和 SARGAN [23]在残差块中引入了终极跳跃连接和空间注意力机制，以更好地保留面部细节。尽管终极跳跃连接具有诸多优点，但有时无法有效地引发表情。

当测试图像与训练时使用的图像相似时，所有这些基于 GAN 的方法都能取得良好效果。然而，将其应用于从真实场景中采集的测试图像时，性能可能会下降。这是因为每个训练数据集都有其自身的偏置，而 GAN 也会学习这种偏置 [24], [25]。这种数据集偏置限制了生成对抗网络的泛化能力。

基于 StyleGAN 的 [6], [7] 潜代码操控方法也已被用于人脸编辑 [26], [27], [28]。首先将输入图像反演到 StyleGAN 的潜空间中，然后编辑潜代码以操控人脸。[29], [27], [30] 中的作者提出了反演技术，以提升输入图像的重建效果并保留精细的面部细节。E4e[29] 学习一个编码器，在重建保真度与可编辑性之间取得平衡。StyleTransformer [27]利用 Transformer 架构提升潜空间反演质量。HyperStyle[30] 采用超网络来预测针对图像的 StyleGAN 生成器适配参数，从而兼顾保真度与可编辑性。StyleRes [31] 引入基于残差的特征变换，以便在反演过程中更好地捕捉精细细节。StyleFeatureEditor (SFE) [32] 专注于细节丰富的反演，通过优化特征表示，在保留人脸属性的同时实现高质量编辑。然而，这些方法仍存在保真度与可编辑性之间的权衡、超出训练域时泛化能力有限以及计算成本高等问题。

近年来，基于回归的变体 [18], [33], [34] 已用于合成逼真表情。Khan等人[18] 提出了掩码回归（MR）网络，其中每个输出像素都会观察输入图像中的一个局部图像块。Akram 和 Khan [33] 通过提出基于像素的岭回归扩展了 MR，其中每个输出像素只关注一个输入像素。尽管这些回归方法能够生成逼真表情，并且对分布外图像具有较好的泛化能力，但往往会产生模糊伪影。在本研究中，我们旨在通过结合基于岭回归的损失与对抗损失来提升生成对抗网络的泛化能力。

RegGAN 的创新之处在于结合了回归方法与基于 GAN 的方法各自的优势。基于 GAN 的方法无法泛化到分布外人脸，而回归方法泛化能力强，但生成的结果较为模糊。RegGAN 引入了基于图像块的回归层以确保泛化能力，并采用基于空间注意力的细化网络。回归损失促使网络学习准确的表情并保留面部细节，而对抗损失则提升真实感和视觉保真度。尽管仅使用少量训练样本，RegGAN 仍能很好地泛化到分布外图像。

这些样本来自 CFEE [35] 数据集，该数据集包含在光照和背景一致的受控环境中拍摄的图像。尽管肖像、虚拟形象、雕像和卡通等图像与训练图像的分布存在显著差异，这种组合仍解决了两项局限，并能在各种人脸图像上实现稳健的表情合成。

III. 提出的回归生成对抗网络（RegGAN）

本节将介绍我们提出的 RegGAN 模型。我们首先概述生成对抗网络框架，然后说明构成该模型的各个不同组件。

A. 生成对抗网络

生成对抗网络（GAN） [36] 旨在学习如何从给定的数据分布中生成样本。GAN 并不显式地对分布进行建模，而是直接学习生成样本。对抗学习框架促使生成的样本更接近真实数据所在的流形。

生成对抗网络由两个相互竞争的网络组成：生成器G和判别器 D。生成对抗网络的训练目标函数定义为

$$\mathcal{L}(\theta_G,\theta_D) = \mathbb{E}_{\mathbf{y}}\left[\log(D(\mathbf{y}))\right] + \mathbb{E}_{\mathbf{z}}\left[\log(1-D(G(\mathbf{z})))\right], \quad (1)$$

其中，$\mathbf{y}$ 是真实数据分布中的一个样本，$\mathbf{z}$ 是随机噪声，θ_D 和 θ_G 分别是判别器D 和生成器 G的参数。以输入图像和/或标签为条件对 G和 D 进行条件化，可实现对条件分布 [1]的建模。给定条件输入$\mathbf{y}$后，条件生成对抗网络的目标函数定义为

$$\begin{aligned}\mathcal{L}(\theta_G,\theta_D) =& \mathbb{E}_{\mathbf{y}}\left[\log(D(\mathbf{y},\mathbf{x}))\right] + \\ & \mathbb{E}_{\mathbf{z},\mathbf{x}}\left[\log\log(1-D(G(\mathbf{z},\mathbf{x}),\mathbf{x}))\right].\end{aligned} \quad (2)$$

本文利用这一点在 [9] 中执行图像到图像转换。

B. 回归生成对抗网络（*RegGAN*）

我们在联合生成器—判别器框架中，学习一个基于回归的表情转换器和一个基于对抗学习的细化网络。生成器网络可分为两部分：i）表情层 G_E 和 ii）细化网络 G_R。表情层 G_E 采用稀疏化回归目标进行训练，可学习准确的表情偏差，并生成能够可靠捕捉目标表情的中间图像。这一步有助于在各种图像上实现泛化。随后，细化网络 G_R 将该中间图像转换为最终输出，在保留所需表情的同时生成逼真的图像。图 3 展示了表情层和细化网络各自的作用。

从数学上看，对于输入图像 $\mathbf{x}$，生成器输出可表示为 $\mathbf{x}_E = G(\mathbf{x}) = G_R(G_E(\mathbf{x}))$，对应于对输入样本进行顺序变换。总体损失函数定义为顺序回归损失与对抗损失之和

$$\mathcal{L}(\theta_E,\theta_R,\theta_D) = \mathcal{L}_E^{\text{Reg}}(\theta_E) + \lambda\mathcal{L}_{\text{GAN}}(\theta_R,\theta_D) \quad (3)$$

其中，θ_E、θ_R 和 θ_D 分别是表情层和细化网络的参数。表情层 $G_E(\mathbf{x})$ 的详细信息见第三节-C小节，而 GAN 生成器和判别器的架构及其各自的损失函数将在第三节-D和第三节-E小节中讨论。

最小化式（3）会使表情层 G_E 通过监督学习和对抗学习进行训练，其中 λ 控制二者的相对重要性。监督学习可确保 θ_E^* 能够很好地泛化到分布外图像，如第三节 C 小节所示。训练完 G_E 后，将其输出用作细化网络 G_R 的输入，随后通过对抗学习训练该网络，以生成照片级真实图像。

C. 表情层

表情层充当一个浅层网络，将输入分布中的样本映射到输出分布。这有助于 RegGAN 在数据集外图像上实现良好的泛化。表情层权重可通过掩码回归 [18], 获得，该方法是岭回归的一种稀疏化改进。具体而言，每个像素都通过 $r\times r$ 感受野表示为其局部邻域的线性组合。与对尺寸为 $m\times n$的输入图像进行岭回归时所用的 mn参数相比，这将每个像素的可学习参数数量减少到 r^2 。研究表明，这种稀疏且局部化的感受野有助于避免过拟合并提升分布外泛化能力 [18], [33]。‘令 $\mathbf{x}_p^n \in \mathbb{R}^{r^2}$ 表示训练样本中输入到 p^{th} 单元的向量化形式 $r\times r$ 。令 $t_p^n \in \mathbb{R}^1$为对应的目标值。对于包含 N个输入和目标值对的训练集，我们可以构造设计矩阵 $\mathbf{X}_p \in \mathbb{R}^{N\times r^2}$ 和目标向量 $\mathbf{t}_p \in \mathbb{R}^{N\times 1}$ ，如下所示:

$$\mathbf{X}_p = \begin{bmatrix}\mathbf{x}_p^{1\top}\\ \mathbf{x}_p^{2\top}\\ \vdots\\ \mathbf{x}_p^{N\top}\end{bmatrix} \quad \text{and} \quad \mathbf{t}_p = \begin{bmatrix}t_p^1\\ t_p^2\\ \vdots\\ t_p^N\end{bmatrix} \quad (4)$$

感受野权重$\mathbf{w}_p \in \mathbb{R}^{r^2\times 1}$和偏置 $b_p \in \mathbb{R}^1$ 对于 p^{th} 像素的损失函数可以写为

$$\mathcal{L}_E^{\text{Reg}}(\mathbf{w}_p,b_p) = \|\mathbf{X}_p\mathbf{w}_p + b_p\mathbf{1} - \mathbf{t}_p\|_2^2 + \frac{\lambda_{\text{Reg}}}{2}\left(\|\mathbf{w}_p\|_2^2 + b_p^2\right) \quad (5)$$

其中 $\mathbf{1}\in\{1\}^N$ 是一个全由 1 组成的向量。使该二次损失函数达到全局最小值的参数可以通过闭式解唯一确定，如下所示

$$\begin{bmatrix}\mathbf{w}_p\\ b_p\end{bmatrix} = \begin{bmatrix}\mathbf{X}_p^\top\mathbf{X}_p + \lambda_{\text{Reg}}\mathbf{I} & \mathbf{X}_p^\top\mathbf{1}\\ \mathbf{1}^\top\mathbf{X}_p & \lambda_{\text{Reg}} + N\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_p^\top\mathbf{t}_p\\ \mathbf{1}^\top\mathbf{t}_p\end{bmatrix} \quad (6)$$

对于大小为 $r\times r$的感受野，线性方程组 (6) 的系统矩阵为 (r^2+1)×(r^2+1)，与岭回归中全局感受野的情况相比，该矩阵非常小。对于

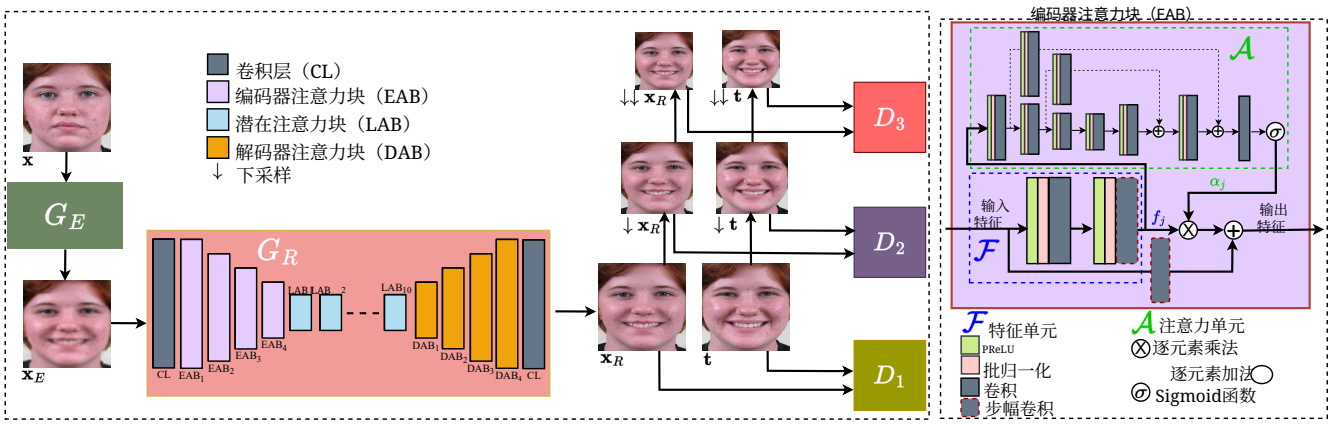

图 2：**左图：**我们的方法由两个部分组成：表情层和细化网络。表情层接收一张输入图像，并生成一张同一人物具有不同面部表情的新图像。随后，细化网络提升合成图像的质量，使其更加逼真、清晰。**右图：**EAB 的架构。



图 3：我们提出的 RegGAN 合成的输入、中间和最终输出示意图。表情层G_E 生成捕捉目标表情的中间图像，而细化网络 G_R 则将这一中间结果转换为逼真的照片级输出。

对于多通道/彩色图像，我们可以通过反复应用 (6)，为每个通道分别学习回归映射。

我们发现，表情层的训练相当高效。对于每个图像到图像转换任务，都需要学习一个不同的表情层。对于表情映射，每对表情都需要训练一个不同的表情层。由于表情层将输入投影到中间表示，因此网络的其余部分无需每次都重新训练。相反，它会将该中间表示投影到输出分布，而不受输入数据分布的影响。这是实现跨输入分布以及跨多个图像到图像转换任务泛化的关键。

D. 细化网络

如图 2 所示，RegGAN 由两个模块组成：基于回归的表情层和基于编码器-解码器的细化网络。细化网络接收中间表示 $\mathbf{x}_E$，将其输入编码器-解码器网络，并生成输出图像 $\mathbf{x}_R$。

细化网络采用注意力块，以在多尺度下恢复$\mathbf{x}_E$ 的面部细节。该网络由三个编码注意力块（EABs）、瓶颈层中的十个潜在注意力块（LABs）以及三个解码注意力块（DABs）组成。

每个注意力块由一个特征单元 $\mathcal{F}$ 和一个注意力单元 $\mathcal{A}$组成。特征单元提取面部特征，而注意力单元则聚焦于眼睛和嘴巴等信息丰富的区域。

注意力单元采用沙漏网络 [37]，随后通过卷积层生成注意力图 $\boldsymbol{\alpha}$。沙漏网络通过在多尺度下结合自底向上和自顶向下的处理方式，旨在捕捉面部区域之间的空间依赖关系。事实证明，这种方法对于人脸超分辨率 [38], [39]和人脸对齐 [40]都很有用。虽然注意力单元在所有块类型（EAB、LAB、DAB）之间共享，但特征单元各不相同。

EAB 采用两个带有 PReLU [41] 激活函数的预激活卷积块（PCB），用于提取面部特征。第二个块使用步幅卷积对特征体进行下采样，在降低计算成本的同时捕获更大的上下文。相比之下，DAB 先采用最近邻上采样扩大特征体，然后再将其传递给卷积块。

j^{th} EAB 的输入特征可表示为 EAB $_{j-1} \in \mathbb{R}^{C_{j-1}\times H_{j-1}\times W_{j-1}}$。 $\mathcal{F}$ 和 $\mathcal{A}$ 的输出通过以下公式计算：

$$\mathbf{f}_j = \mathcal{F}(EAB_{j-1}), \tag{7}$$
$$\boldsymbol{\alpha}_j = \sigma(\mathcal{A}(\mathbf{f}_j)). \tag{8}$$

这里，$\mathbf{f}_j$ 表示特征单元 $\mathcal{F}$ 的输出， $\boldsymbol{\alpha}_j$ 表示注意力单元 $\mathcal{A}$ 的输出，而 σ 是 sigmoid

激活函数。EAB$_{j-1}$ 的输出可写为

$$EAB_j = \Phi(EAB_{j-1}) + \boldsymbol{\alpha}_j \otimes \mathbf{f}_j, \tag{9}$$

其中， $\Phi(\cdot)$ 是步幅卷积层， $\otimes$ 是逐元素乘法运算。

我们采用多种损失函数的组合来训练细化网络。与 [42], 类似，我们采用四种损失：像素损失 $\mathcal{L}_{\text{pix}}$、对抗损失 $\mathcal{L}_{\text{adv}}$、特征匹配损失 $\mathcal{L}_{\text{fm}}$和感知损失 $\mathcal{L}_{\text{per}}$。总体目标是这四项的加权和，定义如下：

$$\mathcal{L}_R = \lambda_{\text{pix}}\mathcal{L}_{\text{pix}} + \lambda_{\text{adv}}\mathcal{L}^G_{\text{adv}} + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}} + \lambda_{\text{per}}\mathcal{L}_{\text{per}}. \tag{10}$$

像素损失： 像素损失的计算方式如下：

$$\mathcal{L}_{\text{pix}} = \frac{1}{m}\sum_{n=1}^{m} \|G_R\left(\mathbf{x}_{E_n}\right) - \mathbf{t}_n\|_1, \tag{11}$$

其中， m 是批次中的训练样本数， G_R 表示细化网络生成器， $\mathbf{x}^{E_n}$ 是 n^{th} 输入图像的中间表示， $\mathbf{t}_n$ 是 n^{th}目标图像， $\|.\|_1$ 表示 ℓ_1 范数。通过最小化这一损失，细化网络能够学习生成高质量的人脸图像，使其与目标图像高度相似。

对抗损失： 为了鼓励生成器生成细节清晰、照片级真实的图像，我们对生成器采用对抗损失。其定义如下：

$$\mathcal{L}^G_{\text{adv}} = \frac{1}{m}\sum_{n=1}^{m}\sum_{k=1}^{3} -D_k\left(G_R(\mathbf{x}_{E_n})\right), \tag{12}$$

其中， D_k 表示尺度为 k的判别器。该损失促使生成器的输出在感知质量方面更接近真实图像。

特征匹配损失： 为稳定 GAN训练并提升生成图像的质量，我们采用特征匹配损失 [43]。该损失最小化中间表示$\mathbf{x}_E$ 的特征图与高质量目标图像$\mathbf{t}$之间的差异。其计算公式为

$$\mathcal{L}_{\text{fm}} = \frac{1}{m}\sum_{n=1}^{m}\sum_{k=1}^{3}\sum_{i=1}^{j}\frac{1}{s^i_k}\|f^i_{D_k}(G_R(\mathbf{x}_{E_n})) - f^i_{D_k}(\mathbf{t}_n)\|_1, \tag{13}$$

其中， j 是 D_k中的层总数， $f^i_{D_k}$ 表示 D_k第 i^{th} 层的特征图， s^i_k 是 $f^i_{D_k}$中的元素总数。

感知损失： 感知损失使用预训练的 VGG-19 网络 [44]，衡量中间表示$\mathbf{x}_E$ 的特征与真实$\mathbf{t}$ 图像之间的感知质量差异。按照式（13）中的记号，感知损失定义为

$$\mathcal{L}_{\text{per}} = \frac{1}{m}\sum_{n=1}^{m}\sum_{i=1}^{l}\frac{1}{s^i}\|f^i_{\text{vgg}}(G_R(\mathbf{x}_{E_n})) - f^i_{\text{vgg}}(\mathbf{t}_n)\|_1. \tag{14}$$

其中， l 表示用于计算感知损失的 VGG-19 层总数， f^i_{vgg} 表示 VGG-19 的第 i^{th} 层的特征图， s^i 表示 f^i_{vgg}中的元素数量。

E. 判别器

我们在细化网络中采用了 3 个判别器 [43], ，分别记为 D_1、 D_2 和D_3，它们具有相同的架构。这些判别器旨在捕捉细粒度人脸细节，以及多尺度下的全局和局部结构。与使用单个判别器相比，这种设计使细化网络能够生成高质量的人脸图像，而这些图像可能会被单个判别器遗漏。多个判别器的有效性已在人脸图像操控 [45]，[46] 和人脸超分辨率 [42], [47]中得到验证。

我们通过分别以 2 倍和 4 倍的因子对输出图像$\mathbf{x}_R$ 和目标图像$\mathbf{t}$进行下采样，构建三尺度图像金字塔。多尺度判别器通常使用的铰链损失 [48], 定义为

$$\mathcal{L}^D_{\text{adv}} = \frac{1}{m}\sum_{n=1}^{m}\sum_{k=1}^{3}[\max(0, 1 - D_k(\mathbf{t}_n) + \max(0, 1 + D_k(\mathbf{x}_{E_n})]. \tag{15}$$

IV. 实验评估

A. 数据集

表情层和细化网络使用复合情绪面部表情（CFEE）[8] 数据集的一个子集进行训练，该子集仅包含七种普遍认可的面部表情：害怕、愤怒、厌恶、中性、快乐、悲伤和惊讶。具体而言，每种表情有 230 张图像，总计 1，610 张图像。我们采用 90/10 的比例划分训练集和测试集。

细化网络最初使用 FFHQ [6]数据集进行训练，该数据集包含 70，000 张高质量人脸图像。所有人脸图像均经过对齐、裁剪，然后调整大小至 $400\times$400。

对于 StarGAN [11] 和 MR [18], 的结果，我们使用 CFEE 数据集训练了模型。GANimation[19], 、SARGAN [23], 、US-GAN [22], 和 DAI2I [21] 的结果则使用预训练模型获得。

为进行定量评估，我们使用了来自 CFEE 数据集的 22张图像，以及从互联网上公开来源收集的22张名人、肖像和头像图像。这些图像可通过 [1]访问。

B. 实现细节

所提出的方法使用 PyTorch 框架实现。训练过程中采用 Adam优化器 [50] , 训练G_R、 D_1、 D_2和 D_3 时使用 $\beta_1 = 0.5$、 $\beta_2 = 0.999$，批量大小设为1。根据 [42], ，我们将式（10）中的损失权重设置为$\lambda_{\text{pix}} = 100$、 $\lambda_{\text{adv}} = 1$、 $\lambda_{\text{fm}} = 10$和 $\lambda_{\text{per}} = 1$。在所有实验中， r的值均设为5。

C. 颜色数据增强

RegGAN 的训练依赖配对数据。具体而言，基于回归的表情层使用中性表情—目标表情图像对进行训练，而细化网络

[1]https://github.com/arbishakram/RegGAN/tree/main/testing-imgs



图 4：RegGAN 在分布外图像上进行面部表情合成的结果，包括厚涂风格人脸（第 1 行）、名人人脸（第 2 行）、幻想图像（第 3 行）、肖像（第 4 行）以及头像（第 5 行）。厚涂风格图像和幻想图像由 Stable Diffusion [49] 生成。所提出的方法在保留输入图像身份和面部细节的同时，生成了令人信服的表情。

则使用表情层的输出及其对应的真实目标表情进行训练。监督模型通常需要大量训练样本才能取得令人满意的结果。为进一步提升训练数据集的多样性并增强表情层的性能，我们对中性表情—目标表情图像对应用颜色增强。这种技术在改变颜色以模拟多种肤色的同时，保留原始人脸图像的形状和几何结构。颜色增强增加了不同肤色样本的多样性，从而提升了回归层的泛化性能。我们使用预训练模型 ReHistoGAN [51], 对图像重新着色。该模型以输入图像和目标图像为输入，生成具有目标图像颜色分布的输出图像。我们为每张图像生成十个增强版本。增强过程的更多细节见补充材料。

D. 评估指标

我们使用四种客观指标对合成图像的质量进行了定量评估：i）表情分类分数，ii）人脸相似度分数，iii）真实感分数，以及 iv）Fr´echet Inception 距离；此外，还通过用户研究进行主观评估。

1）表情分类分数（*ECS*）：我们在 CFEE 数据集上训练了一个 ResNet-18 [52]，用于面部表情识别。随后使用该分类器评估目标表情是否合成正确。ECS 的计算方法是将正确分类的图像数量除以图像总数。

2）人脸相似度分数（*FSS*）：为评估所提出的方法是否保留了面部细节，我们采用人脸相似度分数（FSS）。我们使用 OpenFace [53] 计算合成表情图像与输入图像之间的相似度。

3) 真实感分数（*RS*）：为衡量合成图像的质量，我们使用了 QualiCLIP [54] 分数。CLIP

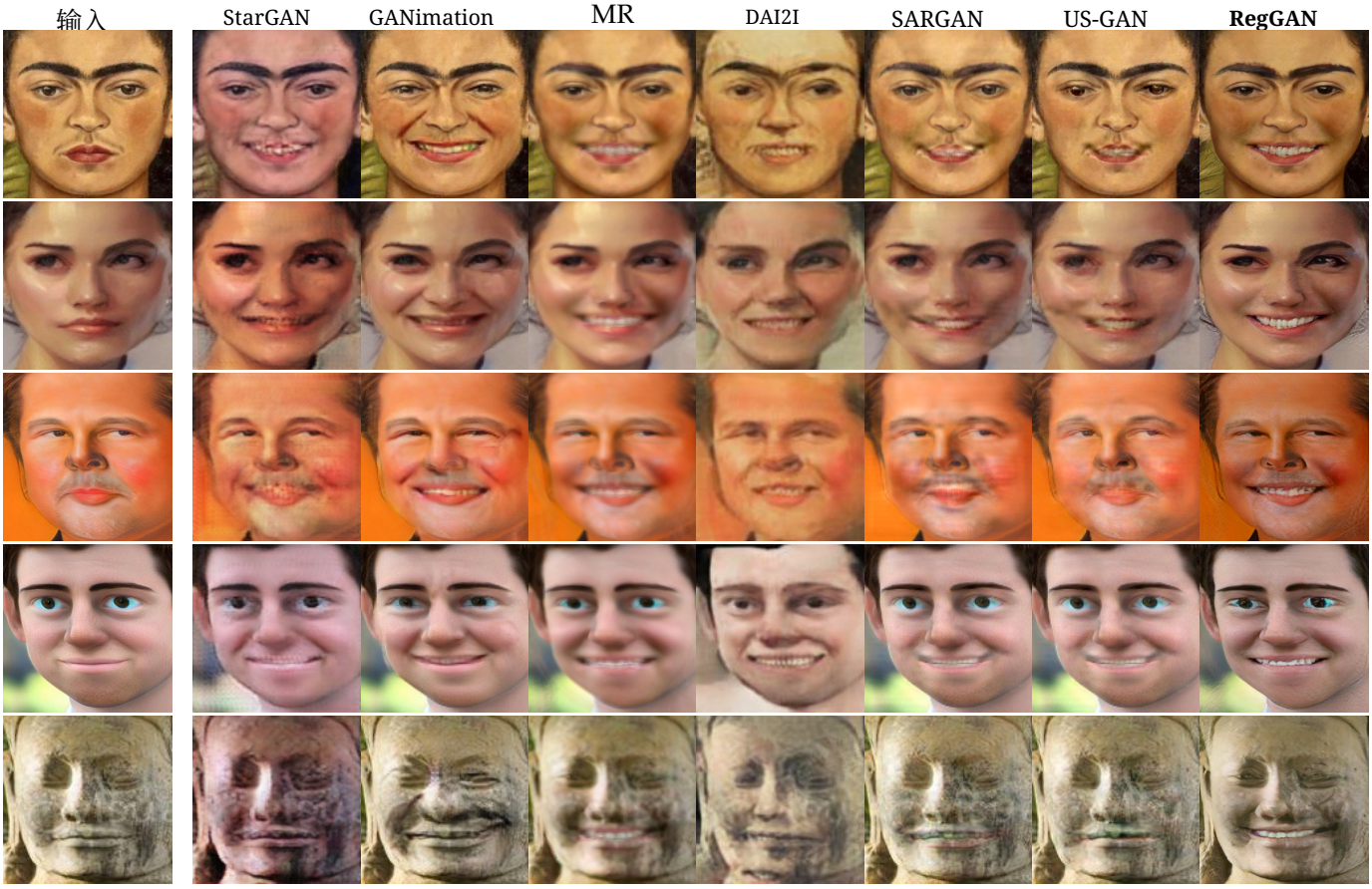

图 5：在分布外人脸图像上，将本文提出的方法 RegGAN 与六种最先进的面部表情合成模型——StarGAN、GANimation、MR、DAI2I、SARGAN 和 US-GAN 进行比较。GANimation 能够合成逼真的表情，但会引入明显的伪影。StarGAN 和 DAI2I 无法在输出中保留输入图像的颜色分布。SARGAN 和 US-GAN 能保留面部细节，但无法生成快乐表情。MR 能生成逼真的快乐表情，但结果往往较为模糊。相比之下，RegGAN 在分布外图像上生成的表情更加清晰、逼真。此外，尽管 RegGAN 仅使用在光照和背景一致的受控环境中采集的 CFEE 数据集进行训练，但它在多种图像类型上都表现稳定，包括从互联网收集的人像、状态图像和虚拟形象。

编码器 [55]为合成图像和七组正负提示词对提取嵌入。提示词对分别为：（“好照片”，“差照片”）、（“清晰图像”，“模糊图像”）、（“清晰边缘”，“模糊边缘”）、（“高分辨率图像”，“低分辨率图像”）、（“无噪声图像”，“有噪声图像”）、（“高质量图像”，“低质量图像”）和（“好图片”，“差图片”）。对于每组提示词对，将余弦相似度通过 Softmax 转换为概率，并选取正向提示词对应的概率。最终分数为所有提示词对对应概率的平均值。

4）弗´雷歇特 *Inception* 距离（*FID*）：为评估生成图像对真实数据分布的近似程度，我们使用了 FID [56]。我们采用预训练的 Inception 网络 [57]提取合成表情图像和真实表情图像的特征。每组特征均以其均值和协方差表示，从而形成一个多元高斯分布。随后计算这两个分布之间的距离作为 FID，距离越小，表示合成图像越接近真实图像。

5）用户研究：为进行主观评估，我们随机选取十张图像，并使用所提方法和六种最先进的基线方法生成相应的合成图像：StarGAN、GANimation、MR、DAI2I、SARGAN 和 US-GAN。图像可通过此链接查看 [2]。对于每张输入图像，都会展示七张带有开心表情的图像，并请参与者回答以下问题：i）哪张图像看起来最真实？ii）哪张图像的表情最令人信服？iii）哪张图像最能保持原始输入图像的身份？图像顺序经过随机打乱，且未透露所用模型。

E. 定性分析

图1展示了 RegGAN 能够生成逼真的表情，并且对未见过的图像具有很强的泛化能力。图4展示了 RegGAN 在六种通用表情上的结果。针对每个输入-目标表情对，都会训练一个独立的回归层。不过，网络的其余部分

[2]https://github.com/arbishakram/RegGAN/tree/main/userstudy-fig

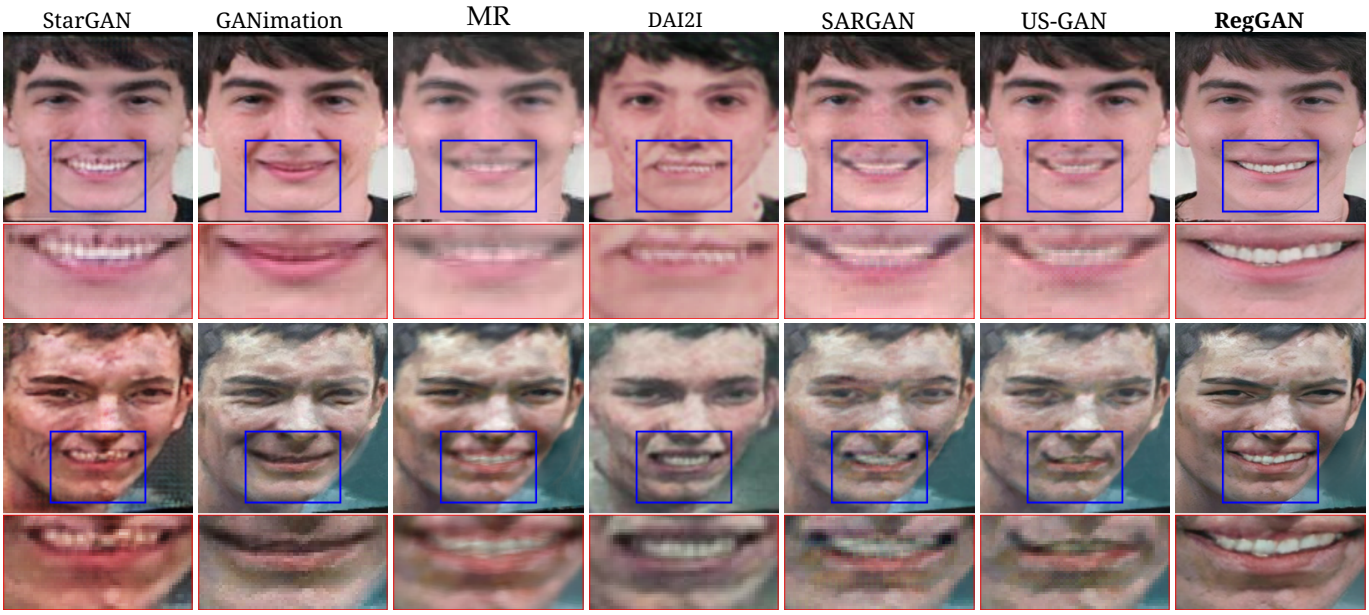

图 6：放大展示了我们提出的方法及其他方法生成的笑容。与所有其他方法相比，我们的方法生成的笑容更加逼真，并增强了单颗牙齿的清晰度，使牙齿之间更易区分、也更清晰可见。

只需训练一次。该实验表明，RegGAN 能够快速泛化到多个表情映射任务。尽管 RegGAN 仅使用人类表情数据集进行训练，却能出色地泛化到肖像、虚拟形象和绘画作品。

图 5 展示了所提出的 RegGAN 与五种最先进方法的比较。测试集包括肖像（第 1 行和第 3 行）、一张厚涂风格人脸（第 2 行）、一张卡通人脸（第 4 行）和一尊雕像（最后一行）。结果表明，GANimation [19] 能够合成逼真的面部表情，但会引入明显的伪影。该方法使用 200,000 张图像进行训练，而我们的方法在应用数据增强后使用 5000 张图像。GANimation 需要精确的面部动作单元提取，而我们的方法无需此步骤。US-GAN [22]和 SARGAN 无法生成令人信服的愉悦表情。相比之下，StarGAN [11] 和 DAI2I [21]难以保留输入图像的颜色和身份细节。基于回归的方法 MR [18],比基于生成对抗网络的方法泛化能力更强，但生成的结果较为模糊。相比之下，所提出的 RegGAN 能够在分布外图像上生成更清晰、更令人信服的表情。这些结果表明，尽管 RegGAN 使用较小的表情合成数据集进行训练，其性能仍然更优。

图6展示了 RegGAN 与竞争方法生成的笑容的放大对比。现有方法通常会生成模糊的笑容，缺乏精确的唇部轮廓，并且无法捕捉牙齿的自然结构。它们通常只是为口腔区域增添白色，而未能捕捉更精细的细节。相比之下，RegGAN 能够生成细节增强的逼真笑容，保留嘴唇的自然曲线，并捕捉牙齿的精细结构，从而实现更清晰的区分。

方法	ECS ↑	FSS ↑	RS ↑	FID ↓
StarGAN [11]	0.71	0.68	0.55	111.9
GANimation [19]	0.45	0.64	0.80	94.09
MR [18]	0.72	0.66	0.19	113.9
DAI2I [21]	0.54	0.64	0.58	140.2
SARGAN [23]	0.61	**0.71**	0.73	98.86
US-GAN [22]	0.66	0.71	0.69	98.92
RegGAN	**0.77**	0.69	**0.84**	**72.48**

表 I：RegGAN 通过合成令人信服、照片级真实的面部表情，在 ECS 和 RS 方面优于其他方法。最低的 FID 表明生成的图像与真实图像的分布高度一致。在 FSS 方面，其取得了第二佳的数值，因为 SARGAN 和 US-GAN 使用终极跳跃连接，能够将输入细节直接复制到输出。

F. 定量分析

在表 I 中，我们对 Reg-GAN 与六种最先进的方法进行了定量比较。基于回归的方法 MR 在保持面部细节的同时，能够生成相当逼真的表情。然而，它在真实感评分和 FID 上的表现有所下降。SARGAN 和 US-GAN 通过终极跳跃连接保持面部细节（FSS），但难以生成逼真的表情。相比之下，RegGAN 在表情合成、身份保持和真实感之间取得了比其他所有方法更好的平衡。图 7 总结了对 40 名参与者进行的用户研究结果。这些结果印证了客观评估指标的结论，显示 RegGAN 在表情、身份保持和真实感方面均优于其他所有方法，获得了最高票数。

G. 消融研究

为了评估潜在注意力块和空间注意力的影响，我们进行了两项消融实验：

1）无潜在注意力块（LABs）：从细化网络中移除所有潜在注意力块。

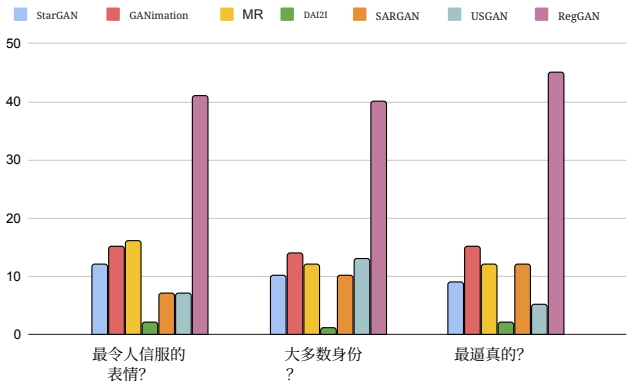

图 7：在合成令人信服且逼真的表情方面，RegGAN 优于所有现有基线方法，同时保留了面部细节。

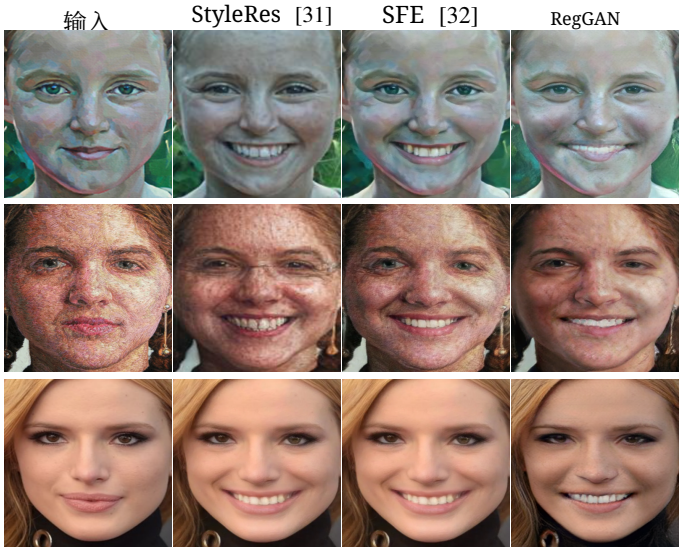

图 8：基于 StyleGAN 的方法能够合成笑容，但往往会丢失精细的面部细节并引入不需要的伪影，例如眼镜（第三行第二列）。所提出的方法在合成逼真的面部表情时，能够更好地保留身份细节。

2）无注意力单元（AUs）：从每个注意力块中移除空间注意力单元。

实验使用 256×256 尺寸的图像进行，批量大小为 3。细化网络首先在 FFHQ 数据集上训练 4 个 epoch，随后在增强后的 CFEE 数据集上进行微调。为确保公平比较，我们使用相同的训练细节，在 256×256 尺寸的图像上对 RegGAN 进行微调。

表 II 展示了这些消融研究的定量结果。加入潜在注意力块和空间注意力单元后，面部细节和表情细节保留效果得到提升。移除 LABs 后，ECS 从 0.85 提高到 0.87，但 FID 升至 80.8，表明分布偏移更大。移除动作单元会导致 FID 和真实感评分最差。完整模型的 FID 最低，且各项指标表现均衡。

方法	ECS ↑	FSS ↑	RS ↑	FID ↓
无 LABs	**0.87**	0.67	0.90	80.8
无动作单元	0.86	0.68	0.88	89.9
RegGAN	0.85	**0.68**	**0.90**	**74.3**

表 II：消融研究的定量结果。引入 LABs 和动作单元有助于在表情、面部细节和真实感之间取得更好的平衡。

方法	ECS ↑	FSS ↑	RS ↑	FID ↓
StyleRES [31]	0.91	0.58	0.28	83.5
SFE [32]	0.70	0.64	0.39	67.4
RegGAN	**0.93**	**0.64**	**0.85**	**61.8**

表 III：RegGAN 和基于 StyleGAN 的方法的定量结果。RegGAN 在表情质量、身份保持和真实感方面得分更高。

H. 与 *StyleGAN* 变体的比较

在图 8 和表 III 中，我们从定性和定量两方面将所提出的方法与基于 StyleGAN 的变体进行比较。虽然基于 StyleGAN 的方法能够生成逼真的面部表情，但难以兼顾表情质量与保真度。其潜在空间操作会将表情与身份纠缠在一起，导致细节缺失和伪影，从而降低真实感。

V. 讨论

基于跳跃连接的方法的局限性：由于表情不会影响面部细节，SARGAN [23] 和 US-GAN [22]采用终极跳跃连接进行面部表情合成。这些方法能够有效地将细节从输入迁移到输出，因此在处理分布外人脸图像时表现出色。然而，尽管它们成功保留了面部细节和颜色细节，却难以生成令人信服的表情。这一局限源于它们直接将输入图像添加到输出中，却未引入任何缩放系数。相比之下，所提出的方法 RegGAN 不仅保留了面部细节和颜色细节，还能生成令人信服的表情，如图 4、5 和 6 所示。

通过回归提升表情泛化：图5和图6显示，基于 GAN 的方法往往难以在多种图像域之间实现泛化，因为它们试图对完整的数据分布进行建模，其中既包括身份变化，也包括表情变化。相比之下，基于回归的方法 MR 专注于从数据中学习表情偏差，使其能够在不同图像中稳定地引入逼真表情。将这两种方法结合到 RegGAN 中，回归可确保表情合成准确，而 GAN 组件则增强真实感。这种混合设计可泛化至人像、虚拟形象、卡通以及雕像（见图4）。如图9所示，RegGAN 在涵盖不同光照、艺术风格和遮挡的多种人脸图像上均具有良好的泛化能力。图10展示了高斯平滑和运动模糊图像上的结果。虽然表情层能够在清晰和模糊的输入上成功生成所需表情，但我们的细化网络产生的锐化效果会减弱原有模糊



图 9：所提出的方法在多种人脸图像上的结果：第 1 行：受到不均匀、多色光照的面孔。第 2 行：采用奇幻艺术家 Peter Mohrbacher 风格的面孔。第 3 行：采用艺术家 Tomer Hanuka 风格的面孔。第 4 行：采用 Gregory Manchess 风格的厚涂画。第 5 行：带有妆容和遮挡的面孔。所有输入图像均使用 Stable Diffusion [58]生成。结果表明，所提出的方法具有很强的泛化能力，能够合成逼真的面部表情，同时保留输入图像的艺术细节与风格。

程度。这在某些情境下可能有益，在其他情境下则并不理想。这种锐化源于一种隐式先验，即输入图像和输出图像都是清晰的。当图像不清晰时，细化网络会强行将其变清晰。未来的工作应着眼于解决保模糊面部表情合成问题。

对配对表情的依赖：尽管 RegGAN 在各种图像上都具有良好的泛化能力，但它依赖由中性表情和目标表情组成的配对数据来训练回归层。当此类数据不可用时，这种依赖可能会限制其应用。未来的工作可以探索非配对或弱监督的替代方案。

细化网络的计算成本：RegGAN 中的细化网络提升了照片级真实感，但也带来了显著的计算开销。表 II 中的结果表明，注意力单元中使用的潜在注意力块和沙漏网络 [37] 能够提升性能，但同时也增加了大量参数，使模型不适用于实时推理或部署在资源受限的设备上。这凸显了一项根本性的权衡：增加深度可以提高真实感，却会降低效率。未来，可以用轻量级替代方案取代注意力单元中的沙漏网络，从而在不牺牲真实感的前提下降低计算成本。

VI. 失败案例

图 11 展示了 RegGAN 在动物面孔上的结果。第 1 行显示，我们提出的方法在猴子脸上合成了逼真的快乐表情。然而，它

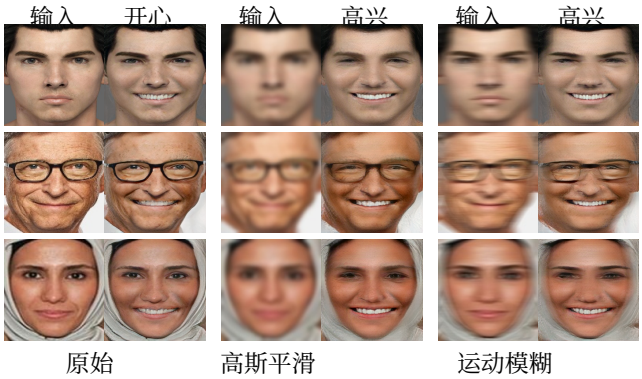

图 10：RegGAN 在高斯平滑图像和运动模糊图像上的结果。表情层能够在清晰和模糊的输入上成功诱导出目标表情，但我们的细化网络产生的锐化效果会使输入

和输出具有不同程度的模糊。这在某些情境下可能有益，但在其他情境下则不那么理想。

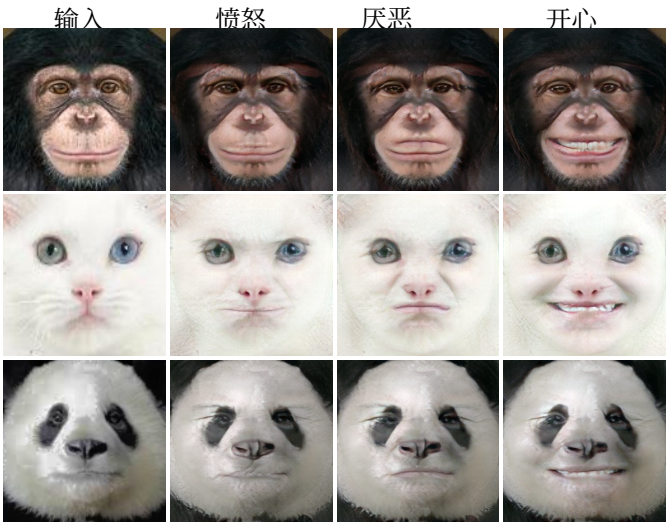

图 11：RegGAN 在猫和熊猫图像中生成类人特征，包括眼睛和面部轮廓。虽然猫天生具有竖瞳，但由于在训练中使用了人脸，RegGAN 往往会将其生成圆形瞳孔。

无法在同一张脸上生成令人信服的愤怒和厌恶表情。第 2 行和第 3 行显示，RegGAN 在猫和熊猫图像上生成了逼真的类人愤怒、厌恶和快乐表情。由于模型是在 人脸上训练的，因此生成的猫和熊猫图像呈现出类人的眼睛和面部轮廓。尽管它能诱导出令人信服的表情并保留面部细节，但拟人化偏置限制了结果的整体质量。

VII. 结论

在这项工作中，我们评估了现有生成对抗网络在分布外图像上的泛化能力，包括 StarGAN、GANimation、DAI2I、SAR-GAN、US-GAN 和 MR。这些生成对抗网络在真实人脸表情数据集上训练，并在训练分布之外的人脸、肖像、虚拟形象、雕像和动物面孔上进行测试。我们的观察

表明，它们对这类图像的泛化能力有限。这一局限源于生成对抗网络试图学习输入图像和输出图像的高维分布，以及它们之间的映射，而训练数据集对这些分布的覆盖较为稀疏。因此，这些生成对抗网络学到的只是对实际分布的估计。结果，当这些生成对抗网络接收到与已学习分布相距甚远的输入图像时，便无法将其映射到目标分布的正确区域。在这项工作中，我们针对这一问题提出了一种解决方案：首先，将输入图像转换为中间表示，然后将该中间表示映射到最终输出分布。第一步转换通过局部感受野学习得到的岭回归层完成，第二步转换则通过优化对抗损失实现。与其他生成对抗网络相比，所提出的 RegGAN 展现出显著更强的泛化能力。

参考文献

[1] Mehdi Mirza 和 Simon Osindero。《条件生成对抗网络》。*arXiv* 预印本 *arXiv:1411.1784*，2014年。[2] Tero Karras、Timo Aila、Samuli Laine 和 Jaakko Lehtinen。《用于提升质量、稳定性和多样性的生成对抗网络渐进式生长》。*arXiv* 预印本 *arXiv:1710.10196*，2017年。[3] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。《去噪扩散概率模型》。33:6840–6851，2020年。[4] Tim Brooks、Aleksander Holynski 和 Alexei A Efros。《Instructpix2pix：学习遵循图像编辑指令》。第18392–18402页，2023年。[5] Manuel Brack、Felix Friedrich、Katharia Kornmeier、Linoy Tsaban、Patrick Schramowski、Kristian Kersting 和 Apolin´ario Passos。《LED- ITS++：使用文本生成图像模型进行无限制图像编辑》。第8861–8870页，2024年。[6] Tero Karras、Samuli Laine 和 Timo Aila。《用于生成对抗网络的基于风格的生成器架构》。载于*IEEE*计算机视觉与模式识别会议，第4401–4410页，2019年。[7] Tero Karras、Miika Aittala、Samuli Laine、Erik H¨ark¨onen、Janne Hell- sten、Jaakko Lehtinen 和 Timo Aila。《无混叠生成对抗网络》。34:852–863，2021年。[8] Shichuan Du、Yong Tao 和 Aleix M Martinez。《复合情绪面部表情》。美国国家科学院院刊，111(15):E1454–E1462，2014年。[9] Phillip Isola、Jun-Yan Zhu、Tinghui Zhou 和 Alexei A Efros。《使用条件对抗网络进行图像到图像转换》。载于*IEEE*计算机视觉与模式识别会议，第1125–1134页，2017年。[10] Jun-Yan Zhu、Taesung Park、Phillip Isola 和 Alexei A Efros。《使用循环一致性对抗网络进行非配对图像到图像转换》。载于*IEEE*国际计算机视觉会议，第2223–2232页，2017年。[11] Yunjey Choi、Minje Choi、Munyoung Kim、Jung-Woo Ha、Sunghun Kim 和 Jaegul Choo。《StarGAN：用于多域图像到图像转换的统一生成对抗网络》。载于*IEEE*计算机视觉与模式识别会议，第8789–8797页，2018年。[12] Taeksoo Kim、Moonsu Cha、Hyunsoo Kim、Jung Kwon Lee 和 Jiwon Kim。《利用生成对抗网络学习发现跨域关系》。第1857–1865页。PMLR，2017年。[13] Rui Huang、Shu Zhang、Tianyu Li 和 Ran He。《超越人脸旋转：用于逼真且保留身份的人脸正面视图合成的全局与局部感知生成对抗网络》。载于*IEEE*计算机视觉与模式识别会议，第2439–2448页，2017年。[14] Xing Di、Vishwanath A Sindagi 和 Vishal M Patel。《GP-GAN：用于从关键点合成人脸的性别保持生成对抗网络》。载于国际模式识别会议，第1079–1084页。IEEE，2018年。[15] Jialu Huang、Liao Jing、Zhifeng Tan 和 Sam Kwong。《多密度草图到图像转换网络》。*IEEE*视频技术汇刊，24:4002–4015，2021年。

[16] Muhammad Mamunur Rashid、Shihao Wu、Yongwei Nie 和 Guiqing Li。《使用基于部件的局部-全局条件生成对抗网络进行高保真面部表情迁移》。《The Visual Computer》，第1–12页，2023年。[17] Yan Zhao、Ziqiang Zheng、Chao Wang、Zhaorui Gu、Min Fu、Zhibin Yu、Haiyong Zheng、Nan Wang 和 Bing Zheng。《利用基于注意力的流水线生成对抗框架进行细粒度面部图像到图像转换》。《Multimedia Tools and Applications》，79:14981–15000，2020年。[18] Nazar Khan、Arbish Akram、Arif Mahmood、Sania Ashraf 和 Kashif Murtaza。《用于学习面部表情合成局部感受野的掩码线性回归》。《International Journal of Computer Vision》，128(5):1433–1454，2020年。[19] Albert Pumarola、Antonio Agudo、Aleix M Martinez、Alberto Sanfeliu 和 Francesc Moreno-Noguer。《GANimation：单次生成解剖结构一致的面部动画》。《International Journal of Computer Vision》，128(3):698–713，2020年。[20] Rongliang Wu、Gongjie Zhang、Shijian Lu 和 Tao Chen。《Cascade EF-GAN：利用局部聚焦渐进式编辑面部表情》。载于 IEEE计算机视觉与模式识别会议，第5021–5030页，2020年。[21] Ying-Cong Chen、Xiaogang Xu 和 Jiaya Jia。《域自适应图像到图像转换》。载于 IEEE计算机视觉与模式识别会议，第5274–5283页，2020年。[22] Arbish Akram 和 Nazar Khan。《US-GAN：终极跳跃连接对面部表情合成的重要性》。《Multimedia Tools and Applications》，第1–1页，2023年。[23] Arbish Akram 和 Nazar Khan。《SARGAN：用于面部表情操控的基于空间注意力的残差》。《IEEE视频技术电路与系统汇刊》，第1–1页，2023年。[24] Luis Herranz、Shuqiang Jiang 和 Xiangyang Li。《使用卷积神经网络进行场景识别：物体、尺度与数据集偏差》。载于 IEEE计算机视觉与模式识别会议论文集，第571–579页，2016年。[25] Antonio Torralba 和 Alexei A Efros。《对数据集偏差的无偏审视》。载于《计算机视觉与模式识别（CVPR），2011年IEEE会议论文集》，第1521–1528页。IEEE，2011年。[26] Tengfei Wang、Yong Zhang、Yanbo Fan、Jue Wang 和 Qifeng Chen。《用于图像属性编辑的高保真生成对抗网络反演》。第11379–11388页，2022年。[27] Xueqi Hu、Qiusheng Huang、Zhengyi Shi、Siyuan Li、Changxin Gao、Li Sun 和 Qingli Li。《用于图像反演与编辑的风格变换器》。第11337–11346页，2022年。[28] Rameen Abdal、Yipeng Qin 和 Peter Wonka。《Image2StyleGAN++：如何编辑嵌入图像？》第8296–8305页，2020年。[29] Omer Tov、Yuval Alaluf、Yotam Nitzan、Or Patashnik 和 Daniel Cohen- Or。《为 StyleGAN 图像操控设计编码器》。《ACM图形学汇刊》（TOG），40(4):1–14，2021年。[30] Yuval Alaluf、Omer Tov、Ron Mokady、Rinon Gal 和 Amit Bermano。《HyperStyle：利用超网络进行 StyleGAN 反演，实现真实图像编辑》。第 18511–18521 页，2022 年。[31] Hamza Pehlivan、Yusuf Dalva 和 Aysegul Dundar。《StyleRes：利用 StyleGAN 转换残差以编辑真实图像》。第1828–1837页，2023年。[32] Denis Bobkov、Vadim Titov、Aibek Alanov 和 Dmitry Vetrov。《细节决定成败：用于细节丰富的 StyleGAN 反演和高质量图像编辑的 StyleFeatureEditor》。第9337–9346页，2024年。[33] Arbish Akram 和 Nazar Khan。《基于像素的面部表情合成》。载于国际模式识别会议，2021年1月。[34] Arbish Akram 和 Nazar Khan。《LSRF：基于全局面部上下文、用于线性面部表情合成的局部化稀疏感受野》。《Multimedia Tools and Applications》，第1–1页，2023年。[35] Shichuan Du、Yong Tao 和 Aleix M Martinez。《复合情绪面部表情》。《美国国家科学院院刊》，111(15):E1454–E1462，2014年。[36] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。《生成对抗网络》。载于《神经信息处理系统进展》，第2672–2680页，2014年。[37] Alejandro Newell、Kaiyu Yang 和 Jia Deng。《用于人体姿态估计的堆叠沙漏网络》。载于欧洲计算机视觉会议，第483–499页。施普林格，2016年。[38] Jonghyun Kim、Gen Li、Inyong Yun、Cheolkon Jung 和 Joongkyu Kim。《用于人脸超分辨率的边缘与身份保持网络》。《Neurocomputing》，446:11–22，2021年。[39] Xin Yu、Basura Fernando、Bernard Ghanem、Fatih Porikli 和 Richard Hartley。《利用面部组件热图引导的人脸超分辨率》。载于欧洲计算机视觉会议

，第 217–233 页，2018。

[40] Adrian Bulat 和 Georgios Tzimiropoulos。我们距离解决二维与三维人脸对齐问题还有多远？（以及一个包含 230,000 个三维人脸关键点的数据集）。载于*IEEE*国际计算机视觉会议，第 1021–1030 页，2017年。[41] 何恺明、张祥雨、任少卿和孙剑。深入探索整流器：超越 ImageNet 分类的人类水平表现。载于*IEEE*计算机视觉与模式识别会议，第 1026–1034 页，2015年。[42] 陈超峰、Dihong Gong、王浩、李志峰和 Kwan-Yee K Wong。学习用于人脸超分辨率的空间注意力。*IEEE*图像处理汇刊，30：1219–1231，2020年。[43] Ting-Chun Wang、Ming-Yu Liu、Jun-Yan Zhu、Andrew Tao、Jan Kautz 和 Bryan Catanzaro。利用条件生成对抗网络进行高分辨率图像合成与语义操控。载于*IEEE*计算机视觉与模式识别会议，第 8798–8807 页，2018。[44] Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。*arXiv*预印本 *arXiv:1409.1556*，2014年。[45] Cheng-Han Lee、Ziwei Liu、Lingyun Wu 和罗平。MaskGAN：实现多样化、交互式人脸图像操控。载于*IEEE*计算机视觉与模式识别会议，第 5549–5558 页，2020年。[46] Yuval Nirkin、Yosi Keller 和 Tal Hassner。FSGAN：主体无关的人脸交换与表情重演。载于*IEEE*计算机视觉与模式识别会议，第 7184–7193 页，2019年。[47] 陈超峰、李晓明、Lingbo Yang、Xianhui Lin、张磊和 Kwan-Yee K Wong。用于盲人脸修复的渐进式语义感知风格转换。载于*IEEE*计算机视觉与模式识别会议，第 11896–11905 页，2021年。[48] Jae Hyun Lim 和 Jong Chul Ye。几何生成对抗网络。*arXiv*预印本 *arXiv:1705.02894*，2017年。[49] Lowfuel。progrock-stable代码库。https://github.com/lowfuel/ progrock-stable，2023年。[50] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。*arXiv*预印本 *arXiv:1412.6980*，2014年。[51] Mahmoud Afifi、Marcus A Brubaker 和 Michael S Brown。HistoGAN：通过颜色直方图控制生成对抗网络生成的图像与真实图像的颜色。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 7941–7950 页，2021年。[52] 何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议，第 770–778 页，2016年。[53] Brandon Amos、Bartosz Ludwiczuk 和 Mahadev Satyanarayanan。OpenFace：一款通用人脸识别库，支持移动应用。技术报告，CMU-CS-16-118，卡内基梅隆大学计算机科学学院，2016年。[54] Lorenzo Agnolucci、Leonardo Galteri 和 Marco Bertini。面向真实世界图像质量评估的质量感知图文对齐。*arXiv*预印本 *arXiv:2403.11176*，2024年。[55] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。第 8748–8763 页。PMLR，2021年。[56] Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。通过双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡。30，2017年。[57] Christian Szegedy、Vincent Vanhoucke、Sergey Ioffe、Jon Shlens 和 Zbigniew Wojna。重新思考用于计算机视觉的 Inception架构。第 2818–2826 页，2016年。[58] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。利用潜在扩散模型进行高分辨率图像合成。载于*IEEE*计算机视觉与模式识别会议，第 10684–10695 页，2022年。