基于 EmoGen 的复现和改进

摘要

本研究报告复现了论文“EmoGen: Emotional Image Content Generation with Text-to-Image Diffusion Models”中的情感图像生成模型并提出改进方案。原论文提出了情感图像内容生成（EICG）任务，通过构建情感空间、映射网络，并结合属性损失和情感置信度等技术，实现了给定情感类别生成高质量情感图像的任务。本报告详细阐述了模型复现过程，包括对原论文的理解、与已有开源代码对比、实验环境搭建、创新点、实验结果分析等内容，并对复现工作进行总结与展望。

关键词：情感图像生成；视觉情感；情感空间；属性损失

1 引言

在当今数字化时代，图像已成为人们表达和交流情感的重要媒介之一。社交媒体平台上，用户频繁分享各类图像以传达自身感受。随着人工智能技术的迅猛发展，计算机视觉领域对图像情感的理解和生成愈发重视。图像生成技术取得了显著进步，能够根据用户输入的文本描述生成逼真的图像。然而，现有文本生成图像模型在处理抽象情感相关图像生成时面临诸多挑战，难以生成具有明确情感且语义丰富的图像。

EmoGen [28] 论文提出了情感图像内容生成（EICG）任务，旨在解决上述问题。该论文通过创新的技术手段，引入情感空间、构建映射网络、设计独特的损失函数等，为生成高质量情感图像提供了新的思路和方法。其技术细节和实验结果在图像生成领域具有重要的参考价值，为进一步研究情感与图像内容生成的关系提供了有力依据。

本研究选题具有重要意义，在理论层面，有助于深入探究情感在图像生成中的作用机制，为计算机视觉和人工智能领域中情感与视觉内容生成的理论研究增添新的维度。通过对情感空间构建和映射的探索，为多模态信息融合提供全新视角，有力推动相关理论体系的发展与完善。在应用方面，复现和深入研究该模型在多个领域展现出广阔的应用前景。例如，在情感化广告设计领域，能够依据目标情感精准生成吸引人的图像，增强广告的感染力与传播效果；于心理治疗辅助场景中，可为患者提供针对性的情感相关视觉刺激，有效辅助治疗进程；在人机交互方面，有助于系统更好地理解用户情感意图，进而生成契合用户情感需求的图像反馈，提升交互体验的质量和效率，具有重要的现实意义和实用价值。

2 相关工作

2.1 视觉情感分析

视觉情感分析是计算机视觉领域的重要研究方向，旨在让计算机像人一样理解图像（或视频）中所包含的情感信息。早期研究主要集中于提取低级特征，如颜色、纹理和风格等。例如，Lee 等 [9] 通过分析颜色图像中颜色与情感的原型关系来评估情感唤醒；Machajdik 等 [12] 提取图像的颜色和纹理等低级特征预测情感类别。随着深度学习技术的兴起，研究人员开始从更高级的语义层次进行情感分析。Borth 等 [1] 提出形容词-名词对（ANP）并构建视觉概念检测器 Sentibank；Rao 等 [17] 构建 MldrNet 从像素、美学和语义层次提取情感线索；Zhang 等 [34] 整合高层内容和低层风格形成更具判别力的情感表示；Yang 等 [29, 31] 从多个对象及对象-场景相关性挖掘情感。然而，现有工作多将视觉情感分析视为分类任务，而本课题关注的是如何生成特定情感的图像，与以往研究方向有所不同。

2.2 文本生成图像

文本生成图像技术旨在将文本描述转化为对应的逼真图像。现有的生成模型主要包括 GANs [5, 10, 35]、VAEs [2, 7, 32]、基于流的模型 [19]、基于能量的模型 [8] 和扩散模型 [20] 等。近年来，扩散模型发展迅速，如 GLIDE [13]、DALLE2 [16]、Imagen [22] 等方法能够生成多样化、逼真且高质量的图像。其中，Stable diffusion [20] 因其稳定的训练和细粒度控制能力而备受关注。针对个性化生成，也涌现出多种基于扩散模型的方法，如 Textual inversion [3] 通过学习新的嵌入实现个性化概念生成；DreamBooth [21] 通过微调网络参数学习新对象。尽管这些模型在生成具体或个性化概念方面表现出色，但在生成抽象情感相关图像时仍存在困难，为本课题的研究提供了改进和拓展的空间。

2.3 图像情感迁移

图像情感迁移 [4] 旨在通过调整图像的颜色和风格等元素来修改其情感基调。早期的图像风格转移研究专注于在不同风格下渲染语义内容，取得了显著的视觉效果。类似地，图像颜色迁移 [18] 致力于调整图像颜色特征。由于颜色和风格对图像情感有重要影响，研究者尝试通过调整这些低级视觉元素来实现图像情感转移。在颜色方面，Yang 和 Peng 等 [27] 首次尝试转移图像颜色；Wang 等 [25] 提出根据给定情感词修改图像颜色的系统，Liu 等 [11] 进一步利用深度学习技术推进该方法；Peng 等 [14] 引入新方法通过引导输入图像的颜色和纹理来改变情感。在风格方面，Sun 等 [23] 和 Weng 等 [26] 在情感感知图像风格转移方面取得了有前景的成果。然而，这些基于颜色和风格调整的方法受固定图像内容限制，导致情感变化不明显，如情感准确率较低（如 29%）[26]。本课题提出的方法基于特定语义生成情感图像，有望克服这一局限性。

3 本文方法

3.1 本文方法概述

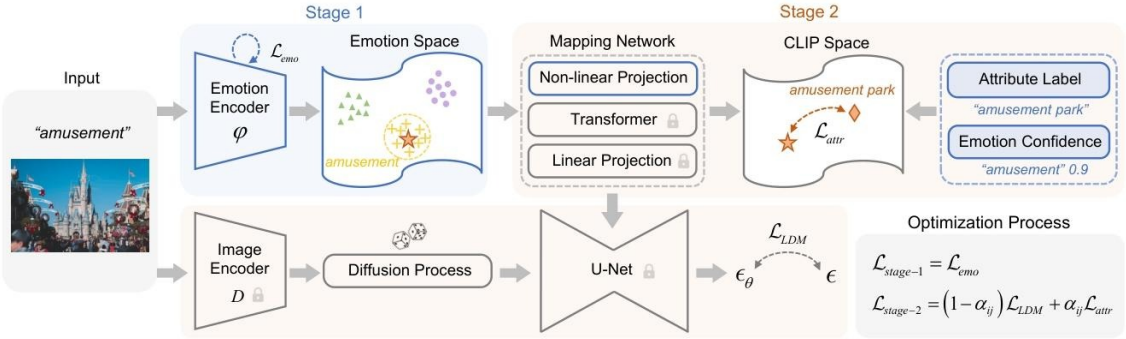


图 1. EmoGen 的训练过程。情感表示（阶段 1）学习一个情感可分的情感空间，情感内容生成（阶段 2）将该空间映射到 CLIP 空间，旨在生成具有情感保真度、语义清晰度和多样性的图像内容。

EmoGen 中提出的情感图像内容生成模型，主要包含情感空间构建、映射网络以及损失函数设计等关键部分，整体框架如图 1 所示。通过这些组件的协同工作，模型能够在给定情感类别时生成语义清晰且情感鲜明的图像内容，有效解决了现有文本生成图像模型在处理抽象情感图像生成时面临的问题。

3.2 情感空间

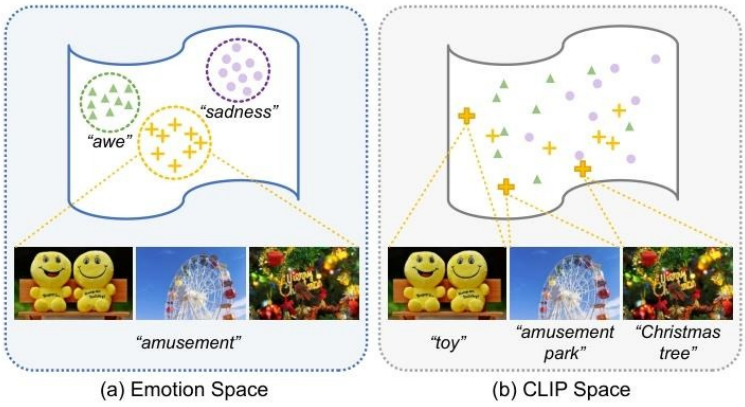


图 2. 尽管 (b) CLIP 空间展示了强大的语义结构，但它很难有效地捕捉 (a) 中所提出的情感空间中的情感关系。

如图2所示，虽然 CLIP 空间有丰富的语义，但是在同一个情感类别下的语义却分得很开，例如 “amusement” 类别下的玩具、摩天轮和圣诞树。为更好地描绘情感关系，EmoGen 引入了情感空间。利用 EmoSet [30] 这个大规模视觉情感数据集，其中每个图像都标注了情感类别，借助 ResNet-50 编码器构建情感空间。在训练过程中，通过交叉熵（CE）损失来训练编

码器，从而使模型能够学习到不同情感的有效表示。损失函数具体为：

$$\mathcal{L}_{emo} = -\sum_{i=1}^{C} y_{emo} \log \frac{\exp(\varphi(x,i))}{\sum_{i=1}^{C} \exp(\varphi(x,i))}, \tag{1}$$

其中 x 表示输入图像，y_{emo} 表示情感标签，C 代表情感类别的总数。一旦损失函数收敛，情感空间得以建立，且在后续的情感内容生成过程中，情感编码器的参数保持固定。在推理阶段，每个情感簇由具有学习到的均值和标准差的高斯分布表示，通过随机从相应高斯分布中采样数据点来获取该情感的表示，这种方式不仅能有效表示情感，还能为情感图像生成过程引入多样性。

3.3 映射网络

即使情感空间在情感上时可分的，但是 Stable diffusion 需要 CLIP 空间中明确的语义作为输入，因此 EmoGen 提出训练一个映射网络，建立情感空间与 CLIP 空间的映射。映射网络利用多层感知机（MLP），并结合非线性操作（RELU），以实现情感空间到 CLIP 空间的映射。具体而言，非线性投影层的结果输入到 CLIP 文本 Transformer 层，进而产生用于 U-Net 的文本嵌入。Transformer 层输出的结束标记嵌入通过线性投影层处理，生成 CLIP 文本特征。特别地，为更好地保留 CLIP 空间中的先验知识，Transform 层和线性投影层的参数保持冻结，仅学习非线性投影层的参数。

3.4 损失函数

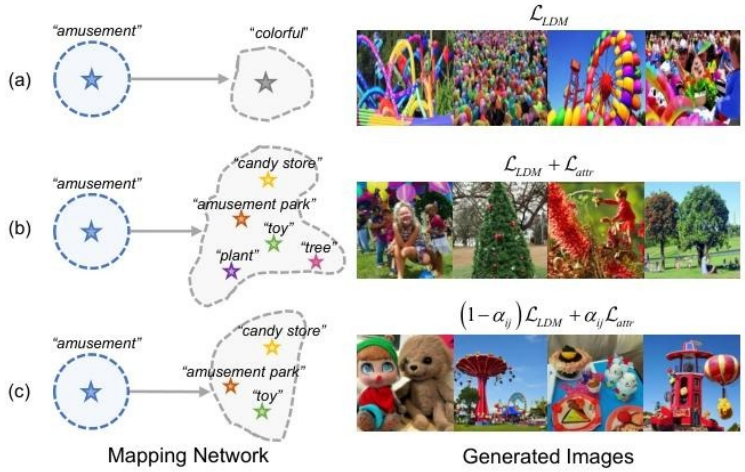


图 3. 损失函数设计的动机。与（a）单独的 LDM 损失相比，（b）属性损失增强了语义清晰度，而（c）情感置信度确保了情感准确性。

3.4.1 扩散损失

现有文本生成图像扩散模型常采用潜扩散模型（LDM）损失 [20] 进行优化，该损失主要用于处理具有单一明确语义的具体实体，然而对于抽象的情感概念，其存在一定局限性。因为每个情感可能包含多种语义，仅使用 LDM 损失可能导致如图3中模型将每个情感收敛到特定语义点，从而失去类内多样性。

$$\mathcal{L}_{LDM} = \mathbb{E}_{z,x,\epsilon,t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, t_\theta\left(F\left(\varphi\left(x\right)\right)\right)\right)\|_2^2\right], \tag{2}$$

其中 ϵ 表示增加的噪声，ϵ_θ 表示去噪网络和 z_t 表示到时间 t 的潜在噪声。

3.4.2 属性损失

为确保生成图像具有清晰多样的语义，利用 EmoSet 数据集中丰富的属性标签（如对象类和场景类型），通过计算生成图像与属性在 CLIP 空间的余弦相似度 [15]，并优化对称交叉熵损失来定义属性损失。

$$\mathcal{L}_{attr} = -\sum_{j=1}^{K} y_{attr} \log \frac{\exp(f(v_{emo}, \tau_\theta(a_j)))}{\sum_{j=1}^{K} \exp(f(v_{emo}, \tau_\theta(a_j)))}, \tag{3}$$

$$f(p, q) = \frac{p \cdot q}{\|p\| \, \|q\|}, \tag{4}$$

其中 a_j 表示属性集合中的第 j 个成员，τ_θ 表示文本编码器，v_{emo} 表示可学习的 CLIP 嵌入，K 表示属性的总数量。这使得每个样本点能够朝着正确语义收敛并远离错误语义，从而在 CLIP 空间中实现每个情感到清晰多样语义的有效映射。

3.4.3 情感置信度

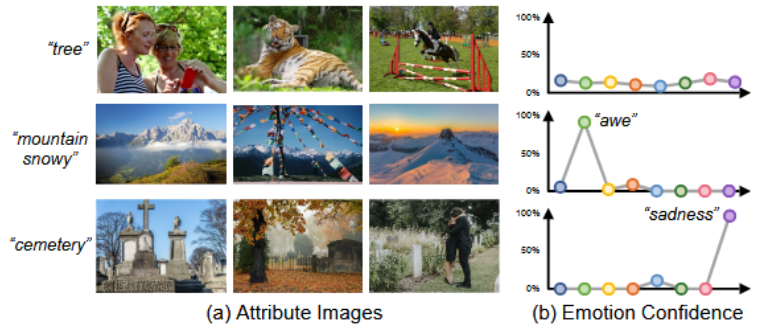


图 4. 情感置信度的例证。每个 (a) 属性由 (b) 八种情绪的置信度分布表示。

考虑到数据集中部分属性可能与情感无关，如图4所示，一棵树可能不会唤醒强烈的情感，但是一只怒吼的老虎可能会令人感到恐惧或愤怒，引入情感置信度来衡量情感与语义属性之间的相关性。通过收集与属性相关的图像并使用预训练情感分类器计算情感分布，为每个属性-情感对分配情感置信度。最终使用情感置信度来平衡扩散损失和属性损失。这样的设计使网络能够根据属性与情感的关联程度，自适应地调整模型的学习重点，从而生成语义明确且情感鲜明的图像。

$$\alpha_{ij} = \frac{1}{N_j} \sum_{n=1}^{N_j} p(x_n, i), \tag{5}$$

其中 x_n 表示输入图像，N_j 表示属性 j 中的总图像数量。

$$\mathcal{L} = \left(1 - \alpha_{ij}\right) \mathcal{L}_{LDM} + \alpha_{ij} \mathcal{L}_{attr}, \tag{6}$$

其中 i 表示情感类别 y_{emo}，j 表示属性类型 y_{att}。情感置信度 $_{ij}$ 越大，表示属性 j 对特定情感 i 的影响越强。低置信度表明属性和情感之间的联系较弱，这表明网络应该从像素级扩散

损失中学习更多信息。当出现更高的置信度时，网络应该优先考虑图像的语义，即属性损失。通过这种设计，EmoGen 有更强的泛化性能，生成语义明确且情感鲜明的图像内容，如图 3 (c) 所示。

4 复现细节

4.1 与已有开源代码对比

在复现的过程中，参考了作者开源的代码（https://github.com/JingyuanYY/EmoGen），并在此基础上添加数据预处理以及模型改进的代码，且该代码可以适配高版本的 diffusers 库。

4.2 实验环境搭建

4.2.1 硬件环境

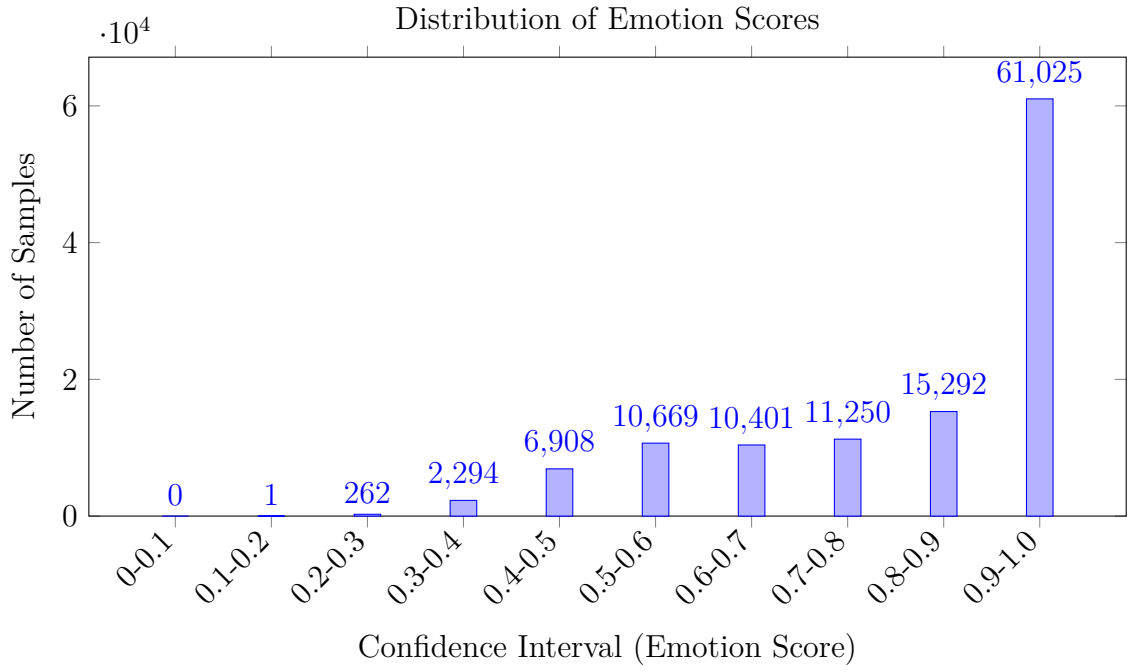
采用一台配备高性能 NVIDIA GPU（如 RTX 4090）的工作站，具备足够的显存（24GB）来处理大规模图像数据和复杂模型计算，以加速模型训练和推理过程。同时，配备了大容量内存（64GB）和高速 CPU（Intel Core i9 - 13900K），确保数据读取和预处理的高效性。

4.2.2 软件环境

操作系统选用 Ubuntu 20.04 LTS，为深度学习框架和相关库提供稳定的运行环境。深度学习框架采用 PyTorch 2.1.1，diffusers 0.21.2。安装 torchvision 库用于图像处理操作；numpy 库用于高效的数值计算；transformers 库用于处理文本数据，包括文本嵌入和分词等操作。此外，还使用了 matplotlib 库进行实验结果的可视化展示。

4.3 创新点

4.3.1 数据预处理



受到图4的启发，为了减少训练成本、让模型学到更能唤醒情感的图像内容，过滤掉情感得分较低的数据。通过反复实验观察结果，结合上面的情感得分的分布图，最后将该数据过滤阈值设置为 0.6。而情感得分的获取则是来源于 CLIP 特征训练的分类器，该分类器的损失函数为：

$$\mathcal{L}_{cls} = -\sum_{i=1}^{C} y_{emo} \log \frac{\exp(\varphi(x,i))}{\sum_{i=1}^{C} \exp(\varphi(x,i))}, \tag{7}$$

其中 x 表示输入图像，y_{emo} 表示情感标签，C 代表情感类别的总数。

4.3.2 微调 CLIP 文本编码器

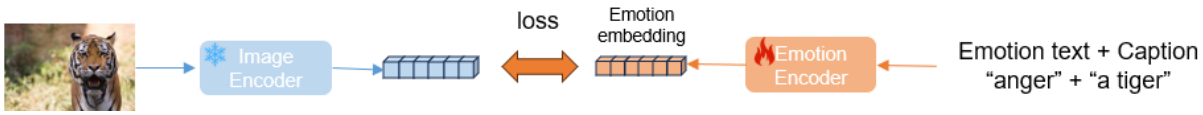


图 5. 微调 CLIP 文本编码器的 Pipeline。

根据图5所示，通过将情感词和对应的图像简述相结合来微调 CLIP 文本编码器，使得 CLIP 文本编码器学习到多元化的情感表达。具体而言，CLIP 是经过大量图文数据对训练得到的，图像中简述的元素对于 CLIP 来说是已知的，但是 EmoSet 中标注的可能对于 CLIP 来说相对陌生，所以可以让模型多到多样化的情感表达。其微调过程的损失函数为：

$$\mathcal{L}_{clip} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{logits}[i,i])}{\sum_{j=1}^{N} \exp(\text{logits}[i,j])} \tag{8}$$

其中 i 和 j 均表示情感和图像简述的连接，N 表示所有情感与所有图像简述进行排列组合的总数量。对于每一行（例如第 i 行），交叉熵损失希望第 i 列的值（正样本）最大化，而其他列（负样本）最小化。该公式中的分母（归一化项）将正样本和负样本的相似度联合建模，目的是在微调过程中尽量少丢失 CLIP 空间原有的先验知识。

5 实验结果分析

5.1 实验设置

数据集：本文所有实验均使用 EmoSet [30] 数据集进行训练，构建情感空间时，按 70:30 比例划分训练集和测试集，训练映射网络时全部数据都作为训练集。而在改进尝试中，按照情感得分的阈值 0.6 进行过滤。

参数设置：采用 AdamW 优化器，初始学习率设为 0.0002，每训练 2 个轮次学习率衰减为原来的 0.8 倍。训练共进行 10 个轮次，每 2 个轮次在测试集评估模型性能，记录 FID、LPIPS、Emo-A、Sem-C 和 Sem-D 指标变化。

评价指标：为了综合评估不同方法在 EICG 任务上的性能，EmoGen 利用常用的指标（FID、LPIPS）并设计一些特定的指标（Emo-A、Sem-C、Sem-D）。1）FID：Frechet Inception Distance（FID）[6] 量化生成图像和真实图像之间的分布距离，提供图像保真度的估计。2）LPIPS：与 [24] 类似，EmoGen 采用 LPIPS [33] 来评估整体图像多样性，值越高表示性能越

好。3）EmoA：由于 EICG 旨在创建唤起情感的图像，因此 EmoGen 设计情感准确性来评估目标情感和生成图像之间的情感一致性。4）Sem-C：人们在可识别的内容下很容易唤起情绪。因此，EmoGen 引入语义清晰度来评估生成的图像内容的明确性。5) Sem-D：情绪是复杂的，每种情绪都可以由多种因素触发。为了覆盖各种潜在场景或对象，EmoGen 推导语义多样性来估计与每种情感相关的内容丰富度。

5.2 定性评估

图 6. 复现结果展示

从生成图像（如图 6 所示）来看，复现的图像结果质量总体不如原文所展示的结果，仅在部分结果中也能看到较为清晰的语义，例如“amusement”中的游乐园旋转木马，“anger”中的游行示威、火灾，“sadness”中的墓碑等。而在“awe”中出现颜色比较奇怪的天空，“disgust”中较为杂乱的图像内容。

图 7. 改进结果展示

从生成图像（如图 7 所示）来看，改进后的图像结果质量总体比复现的结果要好，尽管还是不如原文所展示的结果，但例如“anger”中的怒吼的老虎，“fear”中的饰演鬼的装扮，“excitement”中的室内舞蹈活动等都比复现的结果更能带来情感。

5.3 定量评估

如表1所示，是本次课程作业所有的实验结果。

表 1. Comparisons with the state-of-the-art methods and ablation studies on emotion generation task, involving five metrics.

Method	FID↓	LPIPS↑	Emo-A↑	Sem-C↑	Sem-D↑
Stable Diffusion [20]	44.05	0.687	70.77%	0.608	0.0199
Textual Inversion [3]	50.51	0.702	74.87%	0.605	0.0282
DreamBooth [21]	46.89	0.661	70.50%	0.614	0.0178
EmoGen [28]	41.60	0.717	76.25%	0.633	0.0335
复现	48.89	0.711	74.12%	0.238	0.0257
改进	45.90	0.719	76.99%	0.347	0.0392

EmoGen 中提出的情感空间、映射网络和属性损失函数，都能使得情感图像内容生成效果优于 TI 和 DB，符合原文的消融实验的结果。在同一实验环境下，改进尝试后比复现的结果都得到了全面的提升。但是无论是复现还是改进的结果，在 FID 和 Sem-C 的指标上都远远落后，对生成图像进行对比、观察，得出以下两点可能存在的因素：

- 对于原文的实现细节没有完全掌握，超参数的调整还未达到最佳。
- 生成图像的数量不足，导致出现一定的偏差，例如生成图像中存在大量纹理图。

情感图像内容生成自身也存在一定的局限性，例如情感并不是由单一的视觉因素而唤醒的，还可能受到颜色、风格等因素，情感与内容之间不是严格的二元关系，其次情感图像内容生成的质量受限于现有的文本生成图像模型。

6 总结与展望

6.1 总结

本次课程成功复现并改进了 EmoGen 情感图像生成模型，通过多种优化措施在实验中取得较好结果，验证了改进方法的有效性，提升了模型性能。但仍存在不足，包括复现结果中存在大量的图像质量较差、单个指标远低于原论文。改进后的模型训练时间仍较长，尤其在增加数据增强和拓展属性标签后，计算量增大，需要进一步优化代码和采用更高效的训练策略来提高训练效率。对于一些复杂情感和情感细微差别，模型理解和表达不够精准，需要深入研究情感语义表示和模型架构，提高对情感的理解能力。

6.2 展望

探索多模态融合：结合音频、文本描述等多模态信息，进一步丰富情感图像生成内容，提升生成图像的情感表现力。

优化模型架构：研究更适合情感图像生成的网络架构，如引入 Transformer 结构改进语义理解和信息传递，提高生成图像质量。

拓展应用领域：将模型应用于虚拟现实、智能教育等更多领域，探索其在不同场景下的应用潜力，为实际应用提供更多支持。

参考文献

[1] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, page 223–232, New York, NY, USA, 2013. Association for Computing Machinery.

[2] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *ArXiv*, abs/2203.13131, 2022.

[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

[4] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, October 2020.

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc.

[7] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[8] Yann LeCun, Sumit Chopra, Raia Hadsell, Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. 2006.

[9] Joonwhoan Lee and EunJong Park. Fuzzy similarity-based emotional classification of color images. *IEEE Transactions on Multimedia*, 13(5):1031–1039, 2011.

[10] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18166–18175, 2022.

[11] Dang Jing Liu, Yaxi Jiang, Min Pei, and Shiguang Liu. Emotional image color transfer via deep learning. *Pattern Recognit. Lett.*, 110:16–22, 2018.

[12] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 83–92, New York, NY, USA, 2010. Association for Computing Machinery.

[13] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2021.

[14] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868, 2015.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[16] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[17] Tianrong Rao, Min Xu, and Dong Xu. Learning multi-level deep representations for image emotion classification. *Neural Processing Letters*, 51:2043 – 2061, 2016.

[18] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[19] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1530–1538. JMLR.org, 2015.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[21] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023.

[22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc.

[23] Shikun Sun, Jia Jia, Haozhe Wu, Zijie Ye, and Junliang Xing. Msnet: A deep architecture using multi-sentiment semantics for sentiment-aware image style transfer. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[24] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua, and Nenghai Yu. Diverse semantic image synthesis via probability distribution modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7962–7971, 2021.

[25] Xiaohui Wang, Jia Jia, and Lianhong Cai. Affective image adjustment with a single word. *Vis. Comput.*, 29(11):1121–1133, November 2013.

[26] Shuchen Weng, Peixuan Zhang, Zheng Chang, Xinlong Wang, Si Li, and Boxin Shi. Affective image filter: Reflecting emotions from text to images. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10776–10785, 2023.

[27] Chuan-Kai Yang and Li-Kai Peng. Automatic mood-transferring between color images. *IEEE Computer Graphics and Applications*, 28(2):52–61, 2008.

[28] Jingyuan Yang, Jiawei Feng, and Hui Huang. Emogen: Emotional image content generation with text-to-image diffusion models. *arXiv preprint arXiv:2401.04608*, 2024.

[29] Jingyuan Yang, Xinbo Gao, Leida Li, Xiumei Wang, and Jinshan Ding. Solver: Scene-object interrelated visual emotion reasoning network. *IEEE Transactions on Image Processing*, 30:8686–8701, 2021.

[30] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *ICCV*, 2023.

[31] Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. Stimuli-aware visual emotion analysis. *IEEE Transactions on Image Processing*, 30:7432–7445, 2021.

[32] Chenrui Zhang and Yuxin Peng. Stacking vae and gan for context-aware text-to-image generation. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–5, 2018.

[33] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[34] Wei Zhang, Xuanyu He, and Weizhi Lu. Exploring discriminative representations for image emotion recognition with cnns. *IEEE Transactions on Multimedia*, 22(2):515–523, 2020.

[35] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5795–5803, 2019.