基于情感分类器的情感分解

摘要

近年来在文生图生成任务领域有很大的进展，用户可以用自己的语言生成高质量的图像。但是，现有的文生图的模型在生成给定的物体表现十分好，但是在生成抽象的概念还是有局限性，像是生成情感相关的图像，更多的是局限在具体的物体上。在现实中，用户分享的图像或者是摄影作品一般都不是固定物体的，但是都是传递特定的情感或者感受。在这篇论文里，我提出了 Emotion decomposition，一个利用了 prior 技术，将 CLIP 空间里文本空间和图像空间联系起来的技术，通过不断迭代寻找符合对应情感的 embedding 点，从而生成各种语义的图像但具有同一情感表达的图像。我的方法不仅仅可以应用在情感领域上，也可以应用在别的抽象概念领域上，只要拥有对应的分类器。我还对生成的图像进行了分析，找出了跟情感相关的语义元素，对情感这一领域提供了解释性的路径。

关键词：扩散模型；文生图模型；视觉情感分析

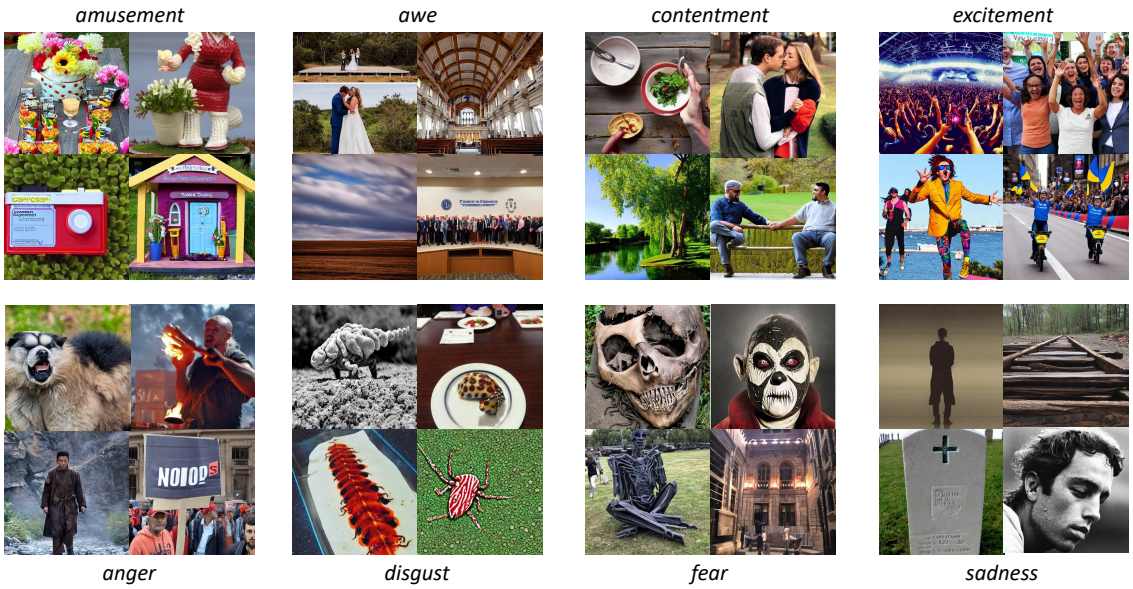

图 1. Emotion Decomposition。我的网络框架可以自主找到情感相关的元素，并生成有明确语义的，表达对应情感的且多样的图像。

1 引言

情感是极具感染力的，鼓舞人心，发泄压力，表达内心世界有力的方式。随着社交媒体的发展，越来越多用户通过细心挑选的图像来分享自己的情感，试图让看到图像的用户能够感受

到他们的情感。为了探索人们看到图像时所产生的情感，视觉情感分析任务（Visual Emotion Analysis）[25,36,38] 这一具有挑战性的任务在计算机视觉领域应运而生。近些年这个领域的发展给市场广告 [5] 和心理健康 [14]，建议挖掘 [35] 带来许多有潜力的应用。

得益于 diffusion 模型的到来和发展 [6,13,28]，文生图的模型得到了前所未有的进展，用户可以通过描述自己想要的具体图像来获得高质量且符合要求的图像 [10,29,40]。现存的文生图模型在生成具体物体时表现的非常出色，但是在给予抽象的概念时表现有所欠缺，一般体现在语义多样性的单一和图像的崩坏上。在现实生活上，像是摄像师在想传达复杂情感时，都不是固定一个物体来表达的。比如说游乐园和圣诞树都可以传达愉悦的情感，但是这两个物体在语义上相差甚远，在情感上却拥有一致性。

如何才能让机器理解情感呢？生成情感是一个很有挑战的任务，因为情感是抽象的，是认知层面的概念，而图像是具体的，是感知层面的概念。情感和图像之间的情感鸿沟 [12] 一直都是视觉情感领域专家们在克服的问题。目前有些出色的工作 [22,31,34] 想要从颜色和纹理上对图像进行修改，从而达到生成情感的目的。但是他们 [34] 都面临了一些障碍，就是情感分类准确率低的问题，由于固定图像内容，而这很大程度会影响图像所传达的情感。有心理学研究 [1,3,4] 表明，除了颜色和纹理会影响情感，图像里的语义内容会更大程度的影响人们所唤醒的情感。

在这篇论文里，我将通过 Emotion decomposition 来实现生成具有多样性且情感一致的图像，且通过对生成图像分析出，对情感相关性高的语音内容，为视觉情感领域揭开可理解的情感生成的新道路。那如何将情感和语义联系，将语义和图像联系起来？在这篇论文里我对语义层面进行研究，对情感进行拆解成各个对应的情感语义要素。CLIP（Contrastive Language-Image Pre-training）[23] 是一个强大的大规模的视觉语言模型。我将使用 CLIP 来将语义和图像进行链接。CLIP 空间里分为了文本空间和图像空间，这两个空间相似但不完全重叠。为了使得文本空间和图像空间能够链接到一起，我使用了 [1] 他们提出的 prior 模型进行链接。目前我们就将情感-语义-图像形成了个生成链了，但是我们要怎么确定情感对应的语义呢？通过我提出的 Emotion decomposition 结合在 CLIP 图像空间学习的情感分类器，我们可以不断迭代地学习对应的情感语义，最后实现生成多样性且情感一致的图像。

总的来说，我的贡献就是：

- 我提出了 Emotion decomposition，实现了生成具有多样性且情感一致的图像。
- 这个方法找出了于情感相关的语义，为视觉情感领域提供了可理解的路径。

2 相关工作

2.1 文生图

文生图生成旨在将文本描述转变成对应的图像。现在的生成模型可以分为，对抗生成模型 [11,19,43]，VAE [9,15,39]，基于流的 [27]，基于能量的 [17]，基于扩散模型的 [6,13,28,29,40]。近年来基于扩散模型的生成模型得到来很大的进展和令人惊叹的效果，有些方法比如说 GLIDE [21]，DALLE2 [24]，Imagen [30]，可以生成多样的，真实高质量的图像。值得一提的是，stable diffusion [28] 是最出名的扩散模型之一，得益于它稳定的生成和快速的生成速度，且有一个活跃的社群在维持。至于个性化的生成，有的模型 [7,10] 会训练出一个新的

嵌入向量来表示新的个性化物体，有的模型 [16, 29, 33] 会对模型内的参数进行微调。Textual Inversion [10] 和 DreamArtist [7] 能够使用仅仅几张图就可以生成个性化的物体，而不用微调模型的参数。Custom diffusion [16] 只用更新交叉注意力模型里的 key 和 value 参数就可以实现个性化。至于多样性的生成，DALLE2 可以实现相同语义不同表现的效果。现存的文生图模型都能很好的生成具体的物体，甚至是个性化的物体，但是面对抽象的概念还有所不足。因此生成情感图像依旧是有挑战的关键的任务。

2.2 视觉情感分析

专家在视觉情感分析探索已经两个世纪了，所使用的方法也从早期传统的方法 [2, 18, 20] 变成最近的深度学习技术 [26, 36, 37, 42]。由于视觉情感的复杂性和抽象性，前期专家使用最有影响力的因素来进行分析，从低等级的视觉因素（颜色，纹理和风格） [18, 20, 26, 42] 到高级的视觉因素（语义） [2, 26, 36, 37, 42]。传统的视觉情感分析都是当作分类任务做的，像是当人们看到这张图像感受到的情感。现在我要做的任务是，给定一个情感，生成对应的能让人感受到这个情感的图像。只有我生成了某个语义的图像，我们才能说这个语义对带来的感情起到多大的作用。

3 准备工作

我在 Karlo 模型 [8] 上应用 Emotion decomposition，同时结合潜在扩散模型算法和扩散 prior 模型实现。

3.1 潜在扩散模型

在潜在扩散模型中（LDM），扩散过程在潜在的首先，训练一个编码器 E，将给定的图像 $x \in \chi$ 映射到潜在代码 $z = E(x)$，同时解码器 D 同时负责重建原始输入图像，使得 $D(E(x)) \approx x$。给定自动编码器，一种去噪扩散概率模（DDPM）被训练用于在数据中产生潜在代码。学习到潜在空间。在去噪过程中，扩散模型可以基于额外的输入进行调节训练 DDPM 模型以最小化目标函数，该目标函数由以下公式给出：

$$\mathcal{L} = \mathbb{E}_{z,x,\epsilon,t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, c\right)\|_2^2\right], \tag{1}$$

去噪网络 ϵ 的任务是正确去除在给定 z_t 的情况下，当前噪声 ϵ 被添加到潜在代码 z_t 中时间步长 t 和调节向量 c

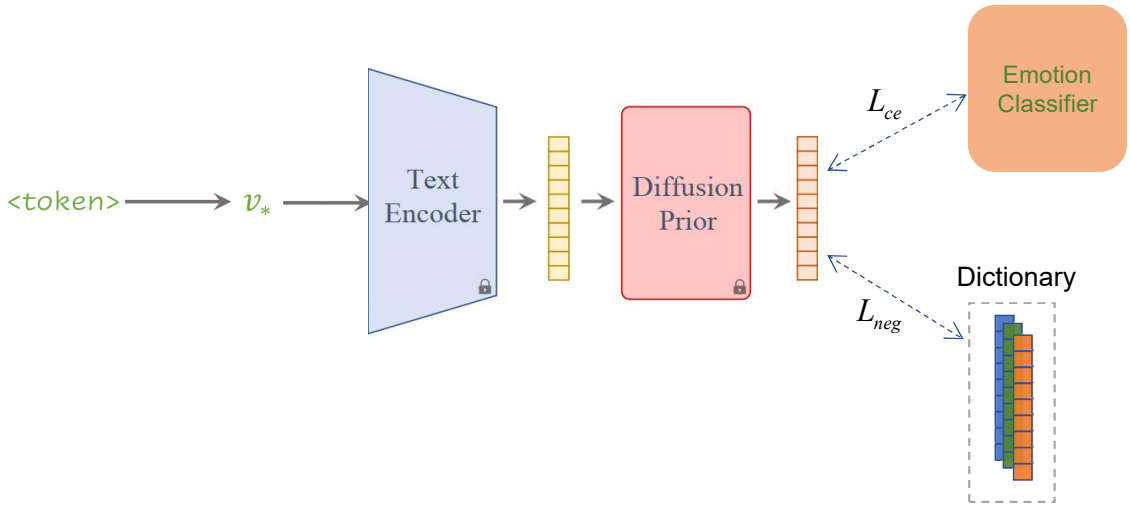

图 2. 方法示意图

3.2 扩散先验

扩散模型通常使用直接从给定文本提示 y 的 CLIP [23] 文本编码中导出的条件向量 c 进行训练。在 Ramesh 等人 [24] 的研究中，他们提出将生成性文本到图像问题分解为两个步骤。首先，使用扩散先验模型从给定的文本提示中预测图像嵌入。接下来，将图像嵌入输入到扩散解码器中，该解码器经过训练，可以根据图像嵌入生成图像。训练通常分为两个独立步骤。扩散解码器使用方程2 中定义的目标进行训练，其中图像嵌入作为条件 c。扩散先验模型 P 随后负责直接从噪声嵌入 et 预测去噪图像嵌入 e:

$$\mathcal{L} = \mathbb{E}_{e,y,t}\left[\|e - P_\theta\left(e_t, t, y\right)\|_2^2\right], \tag{2}$$

一旦对这两个模型进行了训练，每个模型都实现了其目标，就可以将它们组合在一起，创建一个完整的文本到图像的管道。这种两阶段的方法被证明可以提高图像多样性，但更重要的是，从我们的角度来看，它提供了对中间 CLIP 图像嵌入向量的直接访问，并允许在该空间中直接引入约束。

4 方法

Emotion decomposition 是用于解决情感图像生成的任务，在这个任务下我们希望能够生成出情感对应的高质量的图像。和 Textual Inversion 一样，我们通过优化一个新的嵌入向量 v_* 来表示我们在预训练文生图模型的文本条件空间的新的情感元素。由于我们不清楚要生成的情感图像里包含了什么元素，且用现有的数据集图像去优化也是比较困难的，由于同一情感类别下的情感图像里内容丰富且表现多样，比如令人愉悦的图像，可以有多彩的气球，也可以有在广场里开心游玩的小孩。取而代之的是，我引入了一个情感分类器和情感字典来限制我们学到的表征，情感向量表达的是对应的情感类别同时又和我已经找到的情感向量不相同。接下来我会介绍我的优化策略，“prior 约束 “和” 情感约束 “。在训练的过程中，我会逐渐的增加我的约束集，即情感字典，来鼓励学习各种各样的情感元素。我的完整的训练策略展示在了图 2 里。在推理时，我新学到的情感向量会通过添加新的嵌入向量的方式加入到提示词里。

4.1 Prior 约束和情感约束

约束

我将 Prior 约束定义在了 diffusion Prior 输出的空间。Prior 约束使用已经学习到的情感字典里的 embedding 作为负面样本。同时我学习了一个情感分类器约束 Prior 的输出，即 CLIP 图像嵌入向量，让它往相对应的情感靠。比如说，在生成愉悦图像时，情感字典里有游乐园和玩具两个正样本，可以简单定义为 D_{neg}={游乐园，玩具}。

损失函数

具体是怎么将这两个约束加上的，首先我将 v_* 和每个约束的字典词，放入相同的提示词里（比如，Professional high-quality art of a . Photorealistic, 4k, HQ）. 每个句子都可以被编码成 CLIP 文本嵌入向量，这个操作我们简写成 $E_y(c)$。接下来我将 $E_y(v_*)$ 传给 Prior，生成对应的 CLIP 图像嵌入向量，文本提示句子由于经过 Prior 会变成句子的具体样例，而这有利于约束的有效性。比如说愉悦的图像就会具体成游乐园进行约束，进而遍历全部有关愉悦的概念。因此我的损失函数定义为:

$$S(D, v_*) = \mathbb{E}_{d\ D}[< E_y(d), P\left(E_y(v_*)\right) >], \tag{3}$$

$$\mathcal{L}_{emo} = -\sum_{i=1}^{K} y_{emo} \log \frac{\exp\left(C(P(E_y(v_*)))^i\right)}{\sum_{i=1}^{K} \exp\left(C\left(P\left(E_y(v_*)^i\right)\right)^i\right)}, \tag{4}$$

$$\mathcal{L} = S(D_{neg}, v_*) + \lambda L_{emo} \tag{5}$$

上面公式，$S(D, v_*)$ 指的是 v_* 对应的 CLIP 图像嵌入向量与情感字典的约束平均的余弦相似度。L_{emo} 是 v_* 对应的 CLIP 图像嵌入向量的情感交叉熵损失;C^i 指的是第 i 个 C 情感分类器分类出的情感得分;P 指的是 Prior 模型，超参 λ 是用来控制两者的平衡。

正则化

当我们的情感字典变大的时候，陷入一个特定类别的惩罚就会极具地变得很小。为了避免这个情况的发生，我引入了额外的损失函数来计算与情感字典约束力最大的相似度:

这个相似度的计算会和方程3结合在一起，通过跟 $S(D, v_*)$ 使得，跟 v_* 最近的约束能受到更大的惩罚。

自适应的情感字典

如果我们能够手动的添加大规模的情感字典，来寻找新的情感元素是最好的。但是手动去定义情感字典是费时费力的，且可能不能准确找到情感相关的元素。所以最后我才用的是自适应的策略，在训练过程中不断的增加情感字段。正如图 3 所示，每当优化到一定步数，我就会将当前的 v* 保存下来到情感字典，作为新的约束。这样子的自适应策略不仅仅可以将当前学习到的元素保存下来，还可以将其从已学到元素移去还没探索的可行的元素，从而达到多样的生成。

表 1. 在情感图像生成任务上与现今方法在 3 个指标上的比较

Method	LPIPS ↑	Emo-A ↑	Sem-D ↑
Stable Diffusion [28]	0.687	70.77%	0.0199
Textual Inversion [10]	0.702	**74.87%**	0.0282
DreamBooth [29]	0.661	70.50%	0.0178
Ours	**0.743**	63.20%	**0.0422**

5 复现细节

5.1 与已有开源代码对比

我使用了 huggingface 开源的 stable unclip 代码，并自行修改了，使得它可以输出中间的嵌入变量，并且使得文本编码器里的嵌入层里的参数可以参与到学习里，此外情感字典和迭代增加字典的方法在原本的 unclip 论文代码里都没有涉及到，且这个生成情感图像的任务也是自己和老师提出的。所以跟已有的开源代码相比，借用了 prior 模型和他们训练好的 stable diffusion 模型及其预训练参数进行图像生成。

5.2 创新点

我的创新点是提出了 Emotion decomposition，实现了生成具有多样性且情感一致的图像并找出了于情感相关的语义，为视觉情感领域提供了可理解的路径。

6 实验结果分析

由于我的方法是第一个应用情感图像生成，我将其对比与最相关的目前最前沿的文生图生成模型：Stable diffusion [28],Textual inversion [10] 和 Dreambooth [29]。Stable diffusion 是通用的图像生成方法，Textual inversion 和 Dreambooth 是专用于个性化图像生成的。

6.1 定性分析

在图 2中，我们可以看到生成图像内容和我们认知的情感对应的内容息息相关，比如说，敬畏这个情感，在 Emotion Decomposition 表现为了婚礼，教堂，自然景观和会议。且生成的图像质量都很高。

6.2 定量分析

为了评估我生成图像的多样性和情感一致性。我使用的指标是 1）LPIPS：跟 [32] 一样，我才用 LPIPS [41] 作为评估我整体图像的多样性，越高的指标表示越好的表现。2）Emo-A: 因为我的任务旨在还是要生成情感图像，所以我使用情感分类准确率来评估我生成的图像和我目标情感的情感一致性。3）Sem-D: 由于情感很复杂，可能被多种因素所唤起，因此生成情感图像时最好能具有多样性，引入这个指标来表示语义的多样。正如表 1所示，我提出的方法在这两个指标上都超过了现有的方法，在 LPIPS 和 Sem-D, 表示多样性的指标上远远超过了它们。全部方法在情感准确率上都达到了差不多的水平，虽然在情感指标上没有超过别的方法，但是它们方法生成的图像更多的集中在一张图像上，导致了它们生成同一个图像就可以

达到较高的情感准确率，而这和这个任务相违背。此外情感分类器一直是这个领域的一个难点，由于它不能很好的预测到人类的情感，这也是以后我们这个领域要解决的难点。

7 总结与展望

总结

在这篇论文里，我提出了 Emotion decomposition 这个生成情感图像的新范式，并通过定性的分析，证明了方法的有效性。方法使用了 prior 模型将 CLIP 文本空间和图像空间联系起来，并通过在 CLIP 图像空间上训练的情感分类器不断迭代填充情感字典去实现寻找未知情感元素的目的。实验结果我的方法表现超过了现有的文生图模型，虽然它们模型并不是用来做这个任务的。

展望

情感能由很多视觉元素所唤起，比如说颜色，风格和内容。在这个论文里，我只考虑了内容对情感的影响。所以之后会考虑到可控生成，控制能对情感影响的元素来进行图像的生成。此外，在这篇论文里，生成的图像偶尔会出现人造的图像或者单纯颜色和纹理组成的图像，之后也需要对这个问题进行解决。

参考文献

[1] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 459–460, 2013.

[2] 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*, pages 223–232, 2013.

[3] Tobias Brosch, Gilles Pourtois, and David Sander. The perception and categorisation of emotional stimuli: A review. *Cognition and Emotion*, 24(3):377–400, 2010.

[4] Linda Camras. Emotion: a psychoevolutionary synthesis, 1980.

[5] Domenico Consoli. A new concept of marketing: The emotional marketing. *BRAND. Broad Research in Accounting, Negotiation, and Distribution*, 1(1):52–59, 2010.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[7] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022.

[8] Jisu Choi Jongmin Kim Minwoo Byeon Woonhyuk Baek Donghoon Lee, Jiseob Kim and Saehoon Kim. Karlo-v1.0.alpha on coyo-100m and cc15m. https://github.com/kakaobrain/karlo, 2022.

[9] 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. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022.

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

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

[12] Alan Hanjalic. Extracting moods from pictures and sounds: Towards truly personalized tv. *IEEE Signal Processing Magazine*, 23(2):90–100, 2006.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[14] Elaine Hsieh and Brenda Nicodemus. Conceptualizing emotion in healthcare interpreting: A normative approach to interpreters' emotion work. *Patient Education and Counseling*, 98(12):1474–1481, 2015.

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.

[17] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1(0), 2006.

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

[19] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18187–18196, 2022.

[20] 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*, pages 83–92, 2010.

[21] 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. *arXiv preprint arXiv:2112.10741*, 2021.

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

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[25] Tianrong Rao, Xiaoxu Li, and Min Xu. Learning multi-level deep representations for image emotion classification. *Neural Processing Letters*, pages 1–19, 2016.

[26] Tianrong Rao, Xiaoxu Li, and Min Xu. Learning multi-level deep representations for image emotion classification. *Neural Processing Letters*, 51:2043–2061, 2020.

[27] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[31] 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*, pages 1–5. IEEE, 2023.

[32] 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.

[33] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023.

[34] Shuchen Weng, Peixuan Zhang, Zheng Chang, Xinlong Wang, Si Li, and Boxin Shi. Affective image filter: Reflecting emotions from text to images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2023.

[35] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33, 2017.

[36] 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.

[37] 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.

[38] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7584–7592, 2018.

[39] 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. IEEE, 2018.

[40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

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

[42] Wei Zhang, Xuanyu He, and Weizhi Lu. Exploring discriminative representations for image emotion recognition with cnns. *IEEE Transactions on Multimedia*, 22(2):515–523, 2019.

[43] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019.