

文章编号: 1006-3080(2025)05-0681-12

DOI: 10.14135/j.cnki.1006-3080.20241110001

基于区域上下文感知增强网络的图像情感迁移

朱继祥, 张 静

(华东理工大学信息科学与工程学院, 上海 200237)

摘要:为实现准确有效的图像情感迁移, 本文提出基于情感建模关系引导的区域上下文感知增强网络。该网络在外部情感知识的指导下联合多种损失函数以实现准确的图像诱发情感迁移; 在该网络中, 提出一个新颖的区域上下文感知块, 通过多种自注意提取图像中不同感受野的上下文特征, 并借助交叉注意力将这些特征进行自适应融合, 更全面地整合图像信息; 在此基础上通过残差连接恢复深度特征融合中丢失的信息, 更准确地保留图像的内容。同时, 提出一种新颖的情感轮引导模块, 该模块基于情感轮中的情感分布, 使用联合损失引导模型准确地迁移图像情感。为了准确有效地评估模型迁移图像情感的能力, 提出情感迁移综合度量, 综合情感类别、情感极性以及情感在情感轮上的位置, 多角度地评估图像情感迁移的效果, 并基于4个风格不同且广泛使用的情感数据集, 构建一个新的数据集FATE。在FATE上进行的大量实验充分验证了提出方法的有效性, 并在情感迁移效果和图像质量上优于其他对比的方法。

关键词:图像情感迁移; 区域上下文感知增强; 情感建模关系; 情感迁移综合度量; 自注意机制

中图分类号:TP391.4

文献标志码:A

图像情感迁移是指在保留图像内容的基础上, 将目标图像的诱发情感迁移至内容图像上, 使内容图像具有与目标图像相同的诱发情感。与图像风格迁移^[1-3]专注于改变图像的视觉风格和艺术表现不同, 图像情感迁移侧重改变图像传达的情感氛围。该研究可以广泛应用于视觉传达领域和心理学与医学交叉领域, 对于情感辅助诊断和人机交互设计具有重要意义。

早期的图像情感迁移研究通过改变图像的一些低层特征^[4-5]来改变图像情感。这类方法根据心理学理论将情感与颜色进行关系建模, 并利用传统的机器学习^[6]方法调整图像的颜色以改变其情感。后来, 一些研究者把图像语义内容与目标情感的相关性引入图像情感迁移中^[7-9], 利用内容图像和目标图像之间的语义匹配度提高情感迁移的准确性。随着深度学习在视觉情感分类任务中取得了巨大成功^[10-12], 研

究人员使用不同类型的深度神经网络提取图像深层特征并将它们融合^[13-15], 用于实现图像情感迁移。

然而, 目前该领域大多数的研究只局限于较为简单的情感颜色关系建模, 或只考虑图像局部的语义与情感的关联性, 而忽略了将人类对情感的先验认知纳入深度学习模型来指导图像情感的迁移。心理学研究表明, 人的过往经历^[16-17]作为一种先验的知识^[18-19], 会影响人对当下所见事物的情感反应。在观众自身情感产生过程中, 其固有的情感知识将发挥引导作用^[20], 对其情感体验产生重要影响。当人观看图像时, 图像传达的视觉信息被传输至大脑进行处理。大脑接收到视觉信息后, 会自动地联系先前储存的情感知识。具体而言, 大脑先产生与相应情感相关联的心理意象^[18], 其依据人类社会所认知的情感建模关系, 例如情感轮^[21], 把心理意象映射至相同或相近的情感, 最终引发相应的情感体验。

收稿日期: 2024-11-10

基金项目: 上海市自然科学基金 (22ZR1418400)

作者简介: 朱继祥 (1999—), 男, 广西人, 硕士生, 主要研究方向为图像情感迁移、图像情感分类。E-mail: y30221050@mail.ecust.edu.cn

通信联系人: 张 静, E-mail: jingzhang@ecust.edu.cn

引用本文: 朱继祥, 张 静. 基于区域上下文感知增强网络的图像情感迁移 [J]. 华东理工大学学报(自然科学版), 2025, 51(5): 681-692.

Citation: ZHU Jixiang, ZHANG Jing. Regional Context Perception Enhanced Network for Image Emotion Transfer[J]. Journal of East China University of Science and Technology, 2025, 51(5): 681-692.

基于上述分析,本文尝试从心理学研究成果出发,提出一个情感建模关系引导的区域上下文感知增强网络 (Emotion Modeling Relationship guided Regional Context Perception enhanced Network, EMR-RCPN), 实现准确地迁移图像情感。在训练阶段,本文提出一个情感轮引导模块 (Emotion-wheel Guided Module, EGM), 该模块利用预训练的情感分类器进行情感预测,以模拟大脑产生心理意象的过程;利用情感轮模拟大脑所认知的情感建模关系作为指导,并联合多种损失函数计算预测情感和目标情感的多角度差距,让模型在训练中感知不同情感的类间差异,以有效地指导模型改变图像情感。

在特征提取阶段,本文将基于 Transformer 的风格迁移技术引入至图像情感迁移中。早期的图像风格迁移工作中^[2-3], VGG^[22] 通常被用作图像的编码器。在构建合适的融合算法和损失函数后,大多数基于 VGG 的风格迁移模型都能将目标图像的风格特征迁移至内容图像。然而, VGG 不能很好地降低图像的内容损失,并且其提取的风格特征缺乏全局的特征分布信息,导致出现内容低保存度和失真问题。An 等^[23] 在其研究中指出,采用 VGG 会导致内容泄露,即在多次图像特征融合的训练过程中,生成图像的内容结构将发生显著变化。因为 VGG 使用的卷积核小,每个卷积核只能感受图像有限的区域。这使得 VGG 难以对图像的全局上下文信息^[24] 进行有效聚合。最近的工作诉诸于基于 Transformer 的风格迁移方法。例如, Deng 等^[1] 和 Zhang 等^[25] 提出基于 Transformer 的编码器^[26] 作为风格迁移模型的特征提取器。此类特征提取器将图像序列化为大小相等的块,通过多头的自注意机制^[27] 对特征序列自适应加权处理,可以在图像的每个通道中实现全局级别的特征融合,有利于对风格纹理及色彩分布进行有效学习。同时,自注意机制直接在特征层面对整张图像的任意两个位置建立联系,更充分地保留整个图像的全局结构信息。由于在图像情感迁移任务中,生成内容保留度高且风格分布合理的图像至关重要,因此,合理地使用并改进基于 Transformer 的编码器结构对图像情感迁移有良好的作用。

本文在 Transformer 编码器的变体——Twins-SVT^[27] 的基础上提出一个新颖的区域上下文感知块 (Regional Context Perception Block, RCPB), 用于增强 Twins-SVT 编码器,并将其作为区域上下文感知增强网络 (Regional Context Perception enhanced Network, RCPN) 的编码器。不同于传统的 Transformer 中使用全局自注意机制建立图像中任意两个块的信息联

系,根据自注意力的范围, RCPB 将图像中的任意块划分为 3 个不同的感受野。具体来说, RCPB 需要提取图像块本身、其所在的纵轴和横轴的区域、以图像块为中心所在正方形区域的上下文特征,并使用交叉注意力融合这 3 种特征。此操作代替了传统 Transformer 中的全局自注意机制,通过划分出不同范围空间的感受野,使得图像中任意块能够感知到更广泛的图像上下文信息,从而提高模型的对图像内容结构的感知能力,以及对图像整体风格的整合能力。

1 模型与算法

本文提出的 EMR-RCPN 框架如图 1 所示。该模型由 3 个主要部分组成: RCPN、EGM 和解码器。在训练和测试阶段的共同部分,内容图像和目标图像被输入至区域上下文感知增强 Twins-SVT 编码器进行特征提取。然后,目标图像特征通过迁移模块与内容图像特征进行融合。最后,解码器将融合特征重建成与内容图像尺寸相同的形状。

为了在训练中有效控制模型迁移图像情感,在所提出的 EGM 中,情感分类器提取新图像的独热标签;同时,目标图像的标签被向量化形成独热标签。EGM 利用获得的标签对,根据情感类别、情感极性以及情感轮上的情感分布,联合多种损失来控制整个模型迁移图像情感。

在 EMR-RCPN 中,本文使用具有 Transformer 结构的 Twins-SVT 编码器提取内容图像和目标图像的特征。在 Twins-SVT 编码器的基础上,本文提出区域上下文感知块进一步增强 Twins-SVT 编码器的全局建模能力。在图像情感迁移的场景中,风格迁移模型不能够自适应地学习图像的特征分布和目标情感之间的复杂关系。因此,本文使用 EGM 作为激励模型,并引入情感分类器以及心理学的研究成果——情感轮^[21] 作为外部知识,在整个训练过程中结合多种不同类型的损失函数指导模型正确地迁移图像的诱发情感。

1.1 区域上下文感知增强网络

区域上下文感知增强网络负责图像特征的提取和融合。它使用两个结构相同的多层编码器分别提取内容图像和目标图像的特征。编码器层的结构包含块嵌入层、区域上下文感知块、残差卷积位置编码生成器和全局子采样注意块。随后,内容图像和目标图像的特征被迁移模块融合。迁移模块^[28] 由一个 3 层结构的 Transformer 解码器组成。每个解码器层包括 3 个组件: LayerNorm 层,多头自注意力层,多层

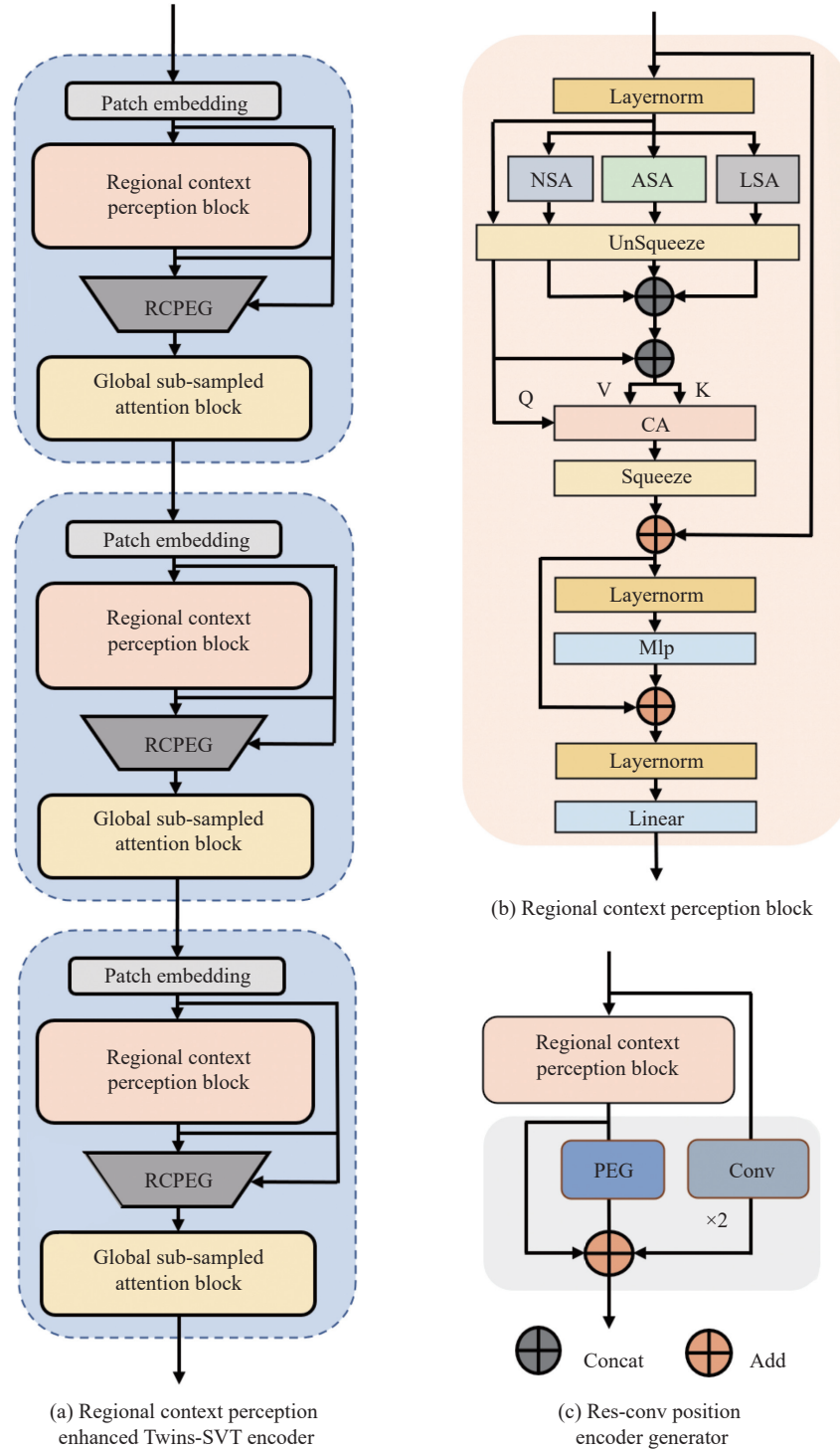


图 2 (a) 区域上下文感知增强 Twins-SVT 编码器结构图; (b) 区域上下文感知块结构图; (c) 残差卷积位置编码生成器结构图

Fig. 2 (a) Regional context perception enhanced Twins-SVT encoder structure diagram; (b) Regional context perception block structure diagram; (c) Res-Conv position encoder generator structure diagram

后的特征送入带有残差连接的前馈网络 (FFN), 增强融合特征的表达能力。最后, 将 FFN 的输出送入 LayerNorm 层和 Linear 层进行调整。具体的公式如式 (7)~(9) 所示:

$$Y^l = \text{Squeeze}(z^l) + z^{l-1} \quad (7)$$

$$z_t^l = \text{FFN}(\text{LayerNorm}(Y^l)) + Y^l \quad (8)$$

$$\hat{z}_{\text{RCPB}}^l = \text{Linear}(\text{LayerNorm}(z_t^l)) \quad (9)$$

其中, Y^l 表示交叉注意力层的输出和区域上下文感知块的输入相加得到的结果, z_t^l 表示带有残差连接的 FFN 输出。通过 LayerNorm 层和线性层的调整后, 模型得到编码器层的输出结果 $\hat{z}_{\text{RCPB}}^l$ 。

特征经过 RCPB 后, 其短距离和长距离的信息被自适应地聚合。为了防止图像特征因过度自注意

而产生退化, 并加强模型处理不同尺寸图像的能力, 引入残差卷积位置编码生成器对特征进行进一步的处理。

1.1.2 残差卷积位置编码生成器 PEG^[27] 根据输入序列的局部邻居信息动态地生成位置编码, 使得 Twins-SVT 能够处理任意长度的输入序列。在实验中, 本文发现引入 PEG 导致训练中出现梯度消失。为了解决这个问题, 本文受到 Conv PE^[29] 和 Twins-SVT^[27] 的启发, 利用卷积层提取位置编码, 并使用残差连接将多个阶段的特征相加。这种结构有助于网络学习到有效的特征表示, 具体来说, 用一个残差连接跨过 RCPB, 并在其中加入两个卷积核大小为 3×3 的卷积块对块嵌入层的输出 z^{l-1} 进行处理。同时, z^{l-1} 也送入 RCPB, 其输出特征 z_2^l 也用一个残差连接跨过 PEG。 z_2^l 送入 PEG 中进行处理。然后将每个模块的输出相加。具体公式如式 (10) 所示:

$$c^l = \text{Conv}_{3 \times 3}(\text{Conv}_{3 \times 3}(z^{l-1})) + \text{RCPB}(z^{l-1}) + \text{PEG}(z_2^l) \quad (10)$$

1.2 情感轮引导模块

本文引入情感轮, 通过更细致地量化目标情感与预测情感之间的差异, 使模型在训练过程中能够

区分迁移错误的样本, 以此提高图像情感迁移的正确性, 并增强模型训练的稳定性。为了模拟人对视觉信息的情感反应, 引入在情感数据集 Flickr^[30] 进行预训练的分类器, 用于预测生成图像的情感。

情感轮引导模块如图 3 所示, 将目标图像的标签和生成图像送入模块中, 生成图像经过情感分类器得到预测标签, 所有标签被张量化成形状为 1×8 的张量。其中, 张量中每个元素的值表示对应情感类别的概率。然后, 计算类别张量对的交叉熵损失, 如式 (11) 所示:

$$L_{ce} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \lg(\hat{y}_{i,c}) \quad (11)$$

在图像情感迁移中, 情感极性间的转换通常更受到关注。因此, 根据情感轮的结构, 本文将 8 种情感分为两种极性, 愉悦、敬畏、满足和兴奋被归为积极情感, 而愤怒、厌恶、恐惧和悲伤被归为消极情感。如果生成样本预测为错误的极性, 将受到严重惩罚。为了得到情感极性的概率, 本文将形状为 1×8 的张量中处于相同极性元素对应的值相加, 得到形状为 1×2 的张量, 其中每个元素表示相应情感极性的概率。计算极性张量对的交叉熵损失, 如式 (12) 所示:

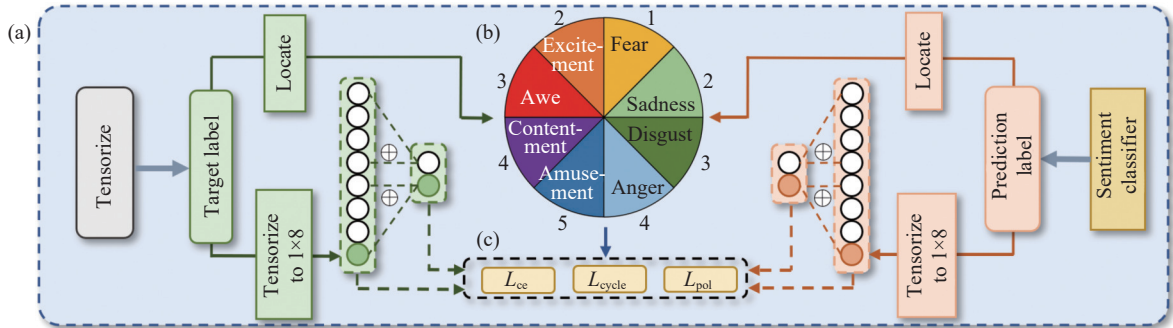


图 3 (a) 情感轮引导模块结构图; (b) 情感轮结构图; (c) 多损失示意图

Fig. 3 (a) Emotion-wheel guided module structure diagram; (b) Emotion wheel structure diagram; (c) Multiple loss diagram

$$L_{pol} = -\sum_{i=1}^N \sum_{c=1}^C p_{i,c}(\text{emo}) \lg(\hat{p}_{emo}(i)) \quad (12)$$

其中, N 代表了一个批次中的样本数量, C 表示情感类别的数量, $p_{i,c}(\text{emo})$ 表示目标图像的真实标签极性编码, $\hat{p}_{emo}(i)$ 表示生成图像的预测标签极性编码。

为了强调预测情感落入不同位置的差别, 本文引入自然对数 e , 把预测情感与目标情感之间的距离作为 e 的指数, 并将其整体作为分子 $e^{d(s,p)}$ 。同时, 以内容图像的情感类别与目标情感之间的距离作为 e 的指数, 将其整体作为分母 $e^{d(s,t)}$ 。情感轮损失被定义为分子与分母的比值。具体计算如式 (13) 所示:

$$L_{cycle} = \frac{1}{N} \sum_{i=1}^N \frac{e^{d(s,p)}}{e^{d(s,t)}} \quad (13)$$

其中, $d(i, j)$ 表示情感轮上情感 i 和情感 j 之间的距离, s 代表起始情感, p 代表预测情感, t 代表目标情感。

使用超参数 α 关联 L_{ce} 、 L_{pol} 和 L_{cycle} , 并将它们的整体定义成情感损失 L_{emo} , 如式 (14) 所示:

$$L_{emo} = (1 - \alpha)(L_{ce} + L_{pol}) + \alpha L_{cycle} \quad (14)$$

最终, EMR-RCPN 的总损失函数如式 (15) 所示:

$$L_{total} = \lambda_s L_s + \lambda_e L_e + \lambda_{id1} L_{id1} + \lambda_{id2} L_{id2} + \lambda_{emo} L_{emo} \quad (15)$$

其中, λ_s 、 λ_e 、 λ_{id1} 、 λ_{id2} 和 λ_{emo} 表示不同损失项的

权重; L_s 、 L_e 、 L_{id1} 和 L_{id2} 表示不同图像的均方误差损失。

1.3 情感迁移综合度量

在图像情感迁移中,在具有情感轮结构的情感类别关系或在 VAD 空间^[31]中进行情感迁移时,生成图像的预测情感可能会落入初始情感和目标情感的区间内。这些样本将被分类至迁移错误的集合中,因此无法对它们进行进一步区分。本文在先前工作^[7,9,15]的基础上,考虑对迁移错误的样本进行区分,提出情感迁移综合度量,更详细且全面地描述模型迁移图像情感的性能。从极性角度来说,落入正确极性的样本获得的惩罚应当比落入错误极性的样本更小;从情感在情感轮的位置来说,如果生成样本的预测情感靠近在目标情感的一定范围内,模型可以容忍这一类型的样本,并根据距离关系给予较轻微的惩罚。

为了展示引入情感极性和情感空间位置关系后,情感迁移综合度量(ETCM)评估图像情感迁移任务的效果,本文通过控制变量的方式,分别示出不同情况下 ETCM 的评估能力,如图 4 所示。表 1 示出了每个示例中各情感类的预测概率。表 2 示出了每

个示例生成样本的情感预测概率、分类得分、极性得分、位置关系得分以及 ETCM。其中,情感极性的的重要性参考 Example 1 和 Example 2 的相关图表,情感变化相对距离的重要性参考 Example 3 和 Example 4 的相关图表。由图 4 和表 1、表 2 可知,图 4(a) 中设置 Example 1 和 Example 2 的目标情感预测概率相同,以及预测情感与目标情感的相对距离相同。由于 Example 2 中生成样本的预测情感与目标情感不在同一类情感极性中,因此其极性得分为 0.35,远低于 Example 1 的 1.85 分。这表示 ETCM 对于极性错误的样本会给予较低的得分,使其对最终 ETCM 的值有较大的影响。由图 4 和表 3、表 4 可知,Example 3 和 Example 4 的目标情感预测概率接近,且情感极性得分相似。由于 Example 3 中生成样本情感变化的相对距离比 Example 4 小,为了增加不同样本的相对距离区分度,在位置关系得分中引入自然常数作为底数,情感移动距离作为其指数。情感变化的相对距离越大,说明样本的预测情感越接近目标情感,其得分相对于远离目标情感的样本越会呈现较大的差异。例如表 2 中,Example 4 的位置变化得分约是 Example 3 的 7 倍。

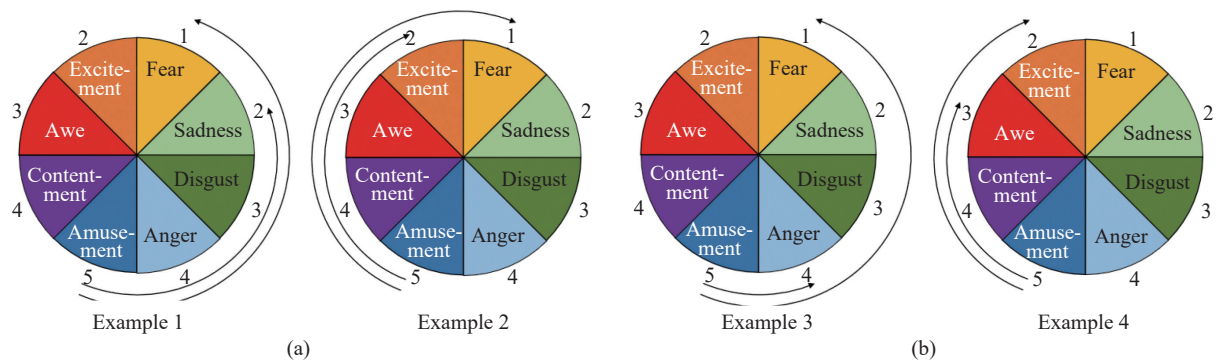


图 4 情感迁移综合度量在 4 个不同示例中的运用

Fig. 4 Application of ETCM in four different examples

表 1 4 个不同示例中各类情感的预测概率

Table 1 Prediction probability of various emotions in four different examples

Example	Prediction probability							
	Fear	Sadness	Disgust	Anger	Amusement	Contentment	Awe	Excitement
1	0.20	0.5	0.05	0.10	0.05	0	0.1	0
2	0.20	0.1	0	0.05	0	0.05	0	0.6
3	0.10	0.1	0.05	0.70	0.05	0	0	0
4	0.05	0.1	0	0.05	0	0.10	0.5	0.2

总之,通过考虑情感类别、情感极性以及情感在情感轮的位置关系来计算 ETCM,可以有效区分被错误分类的样本,更全面地描述模型迁移情感的能力。

具体而言,一个批次的预测标签是一个 $N \times 8$ 维的向量,目标情感在向量中对应的概率值是 p_{ic} 。取 N 个标签的 p_{ic} 均值作为类别得分 s_c ,如式 (16) 所示:

表 2 4 个不同示例中情感、得分以及 ETCM 的详细信息

Table 2 Detailed information on emotion, score and ETCM in four different examples

Example	Target	Source	Prediction	cls_score	pola_score	dist_score	ETCM
1	Fear	Amusement	Sadness	0.20	1.85	0.3679	2.4179
2	Fear	Amusement	Anger	0.20	0.35	0.3679	0.2000
3	Fear	Amusement	Excitement	0.10	1.95	0.0498	2.0098
4	Excitement	Amusement	Awe	0.20	1.80	0.3679	2.3679

表 3 FATE 情感迁移数据集的详细参数

Table 3 Detailed parameter for the FATE emotion

Item	Number of images		
	Content set	Target set	FATE dataset
Total	13 230	646	13 876
Amusement	2017	81	2098
Anger	1041	62	1 103
Awe	2002	82	2084
Contentment	2014	56	2070
Disgust	1290	56	1 346
Excitement	1997	84	2081
Fear	1 149	92	1 241
Sad	1 720	133	1 853

$$s_c = \frac{1}{N} \sum_{i=1}^N p_{i,c} \quad (16)$$

针对情感在情感轮上的位置关系, 计算初始情感与预测情感在情感轮上的矢量距离 $\mathbf{d}_{st}$, 以及其与目标情感的距离 $\mathbf{d}_{sp}$ 。 $\mathbf{d}_{sp}$ 和 $\mathbf{d}_{st}$ 分别作为自然常数 e 的指数, 并计算它们的比率 s_d , 如式 (17) 所示:

$$s_d = \frac{1}{N} \sum_{i=1}^N \frac{e^{\mathbf{d}_{i,sp}}}{e^{\mathbf{d}_{i,st}}} \quad (17)$$

目标情感的极性分数根据预测标签中表示相同极性元素的概率之和计算得出, 称为 s_p 。为了强调极性的重要性, 将样本的 s_p 视为一个关键因素。当样本被分类到错误的极性时, 设置它的 s_p 为 0。相反, 当样本被分类到正确的极性时, 将目标情感所属极性的概率 $p_{i,c}$ (pola) 与 1 相加, 使其值大于 1。然后将新的 s_p 与 s_d 进行求和操作, 如式 (18) 所示:

$$s_p = \begin{cases} \frac{1}{N} \sum_{i=1}^N \sum_{c=1}^5 (1 + p_{i,c}(\text{pola})), & \text{if pola is correct} \\ 0, & \text{else} \end{cases} \quad (18)$$

最后, 将 s_d 、 s_p 和 s_c 按迁移后样本的极性情况进行分类, 获得 ETCM, 如式 (19) 所示:

$$\text{ETCM} = \begin{cases} s_c + s_p + s_d, & \text{if pola is correct} \\ s_c + s_p \times s_d, & \text{else} \end{cases} \quad (19)$$

2 实验设计

为了验证所提出的图像情感迁移方法的有效性, 本文以 4 个广泛使用的情感数据集 FI [32], TwitterLDL [33]、Emotion6 [5] 和 ArtPhoto [34] 为基础, 构建了一个新的数据集 FATE。FI 数据集来源于照片分享平台 Flickr 和 Instagram, 它由 23 308 张图片组成。制作者使用 Amazon Mechanical Turk 将 Mikels 情感轮 [21] 上 8 个情感类别中的 1 种作为每张图像的情感标签。TwitterLDL 数据集来源于 Twitter 网站, 制作者使用各类情感关键词搜索并过滤, 进行图像收集。它共包含 10 045 张图片。其中, 8 名投票者分别对 Mikels 情感轮中的 8 类情感进行投票, 并将投票结果作为每张图片的标签。Emotion6 数据集是基于 Ekman 情感模型 [35] 中的 6 个情感词进行检索而获得的, 其中每个情感类别收集了 330 张图片。ArtPhoto 数据集的图像由专业艺术家拍摄, 共有 806 张图像, 每张图像都标记为 8 类情感中的某一类, 其风格、色彩以及构图与情感的相关性强, 适合作为图像情感迁移任务中目标图像集。

本文在 FATE 数据集上进行了一系列实验, 并复现了当前 4 种有代表性的风格迁移方法, 将其与所提出的方法进行对比。

2.1 FATE 数据集构建及实验设置

本文选择情感轮上的 8 类情感作为情感迁移的类别基础。使用 FI、TwitterLDL 和 Emotion6 作为内容图像集。由于 TwitterLDL 的标签是分布式的, 因此将分布标签中投票数最高的类别作为当前图像标签。

由于面部表情对图像的诱发情感有着显著的影响 [9,36], 因此包含人脸的图像难以进行有效的情感迁移。为了确保实验结果的客观性, 以及更准确地计算图像情感迁移的效果, 本文使用人脸检测器从内容图像集中移除被检测有人脸的图像。由于人脸检

测器的能力有限,只有那些人脸明显且完整的图像被删除。同时,为了使得模型能够更准确地转换不同的情感表达,提高生成图像的情感表现力和真实感,本文对内容图像集进行类平衡操作,帮助模型更全面地学习各种图像特征之间的差异。经过类平衡操作的数据集也可用于图像情感分类任务。此外,本文使用 ArtPhoto 数据集作为目标图像集。FATE 的详细信息如表 3 所示。

本文在一块 NVIDIA GeForce RTX 3090 24GB 显卡上使用 Pytorch 框架训练模型,其初始学习率设定为 0.0001。采用 Adam 优化器和余弦退火学习率调度器,动态地优化模型的参数和学习率。情感分类器是一个 ViT(Vision Transformer)模型^[37],其在 Flickr 数据集^[30]上进行了情感分类微调,并在测试集上达到 0.7048 的准确率。

2.2 对比实验

从 FATE 中随机选择 100 张内容图像和 16 张目标图像作为测试集,将 EMR-RCPN 与 4 种复现的方法进行广泛的对比实验。在评估指标上,选择总分类准确度(Acc)和 ETCM 来评估模型迁移情感的能力。

情感迁移的实验结果如表 4 所示,EMR-RCPN 在总情感分类准确度上的表现优于其他对比模型。其中,EMR-RCPN 的总分类准确率为 24.31%,比 S2WAT^[25]、StyTr2^[11]、Adaattn^[2]和 SANet^[3]平均提高约 4.68%。对于 ETCM,EMR-RCPN 达到 1.2001,获得了最佳结果,对比方法平均提高 0.07038,这表明本文提出的情感轮引导模块对迁移图像情感具有显著的优势。

表 4 情感迁移方面的实验结果

Method	Acc/%	ETCM
SANet ^[3]	14.69	1.0540
Adaattn ^[2]	20.85	1.1585
Stytr2 ^[11]	20.94	1.1239
S2WAT ^[25]	22.03	1.1825
Ours	24.31	1.2001

为了评估模型恢复内容图像内容的能力,本文使用 Reconstructed error^[9]、LPIPS(Learned Perceptual Image Patch Similarity)^[38]、PSNR(Peak Signal to Noise Ratio)^[39]、SSIM(Structural Similarity)^[40]和 FID(Fréchet Inception Distance)^[41]作为指标。生成图像质量的实验结果如表 5 所示,结果表明 EMR-RCPN 实现了最

表 5 图像质量的实验结果

Method	Reconstructed error	LPIPS	PSNR	SSIM	FID
SANet ^[3]	52.25	0.3579	27.98	0.7706	544.13
Adaattn ^[2]	41.54	0.3064	27.97	0.8002	488.49
Stytr2 ^[11]	57.70	0.3875	27.96	0.7450	460.30
S2WAT ^[25]	44.43	0.2887	27.97	0.7960	437.51
Ours	36.32	0.2218	28.02	0.8213	368.07

佳的内容保留效果。与 S2WAT 相比,EMR-RCPN 的内容保留度接近但略高于它,这是因为编码器的区域上下文感知块融合不同区域的上下文特征,聚合丰富的图像信息,并结合残差连接有效地保留了图像内容结构。与 StyTr2 相比,EMR-RCPN 引入了更强大的注意模块,使其保留的图像内容更细致全面。Adaattn 和 SANet 是两种基于 VGG 编码器的图像风格迁移方法。其中 Adaattn 对每个像素进行注意力规范化,以对齐特征分布,但过度规范化会导致风格的退化,使图像失真。SANet 基于图像特征计算注意力图,但其只使用深层的特征来计算注意力得分,忽略了浅层的细节,导致内容信息的丧失。

2.3 消融实验

为了展示区域上下文感知块建模图像全局信息的能力,本文采用原始的 Twins-SVT 编码器作为对比方法,并保持两个方法的解码器相同,实验结果如表 6 所示。与对比方法相比,EMR-RCPN 在 Reconstructed error、LPIPS 和 FID 分别减小 10.72、0.0815 和 86.30, PSNR 和 SSIM 分别增加 0.04 和 0.033。这表明 EMR-RCPN 的编码器在进行情感迁移时保留图像更多的内容信息。

表 6 图像质量的消融实验结果

Method	Reconstructed error	LPIPS	PSNR	SSIM	FID
Twins-SVT+decoder	47.04	0.3033	27.98	0.7883	454.37
Ours	36.32	0.2218	28.02	0.8213	368.07

为了证明情感轮引导模块的有效性,本文从模型中移除该模块,实验结果如表 7 所示。与没有此模块的模型相比,Acc 提高 2.72%,达到 24.31%。ETCM 提高 0.0474。这些结果证明情感轮引导模块在训练中能够发挥指导作用,提高模型迁移图像情感的准确性。

表 8 示出了情感轮引导模块中不同损失函数组合的影响。模型在使用完整的联合损失时达到了最

表 7 图像情感的消融实验结果

Table 7 Experimental results of image sentiment in ablation study

Method	ETCM	Acc/%
RCPN	1.152 7	21.59
EMR-RCPN	1.200 1	24.31

表 8 情感轮引导模块中不同组合损失函数的消融实验结果

Table 8 Ablation experiment results on different combinations of loss functions in the EGM

Loss function	ETCM	Acc/%
RCPN	1.152 7	21.59
RCPN + L_{cycle}	1.166 8	23.61
RCPN + L_{cls} + L_{cycle}	1.194 6	23.88
RCPN + L_{cls} + L_{pola} + L_{cycle}	1.200 1	24.31

佳性能。其他未标注的组合由于梯度消失的问题导致模型训练失败。

2.4 可视化分析

如图 5 所示, 对 EMR-RCPN 与 4 种优秀的风格迁移方法进行了可视化分析。图 5 的第 1 列为内容

图像, 最后 1 列为目标图像, 第 2~6 列示出了不同方法的生成图像。EMR-RCPN 使用区域上下文感知块聚合图像丰富的上下文特征, 使模型能够聚合丰富的图像信息。使用残差连接结构有效地保留了图像的内容, 使生成的图像更符合现实世界的观感。所有比较方法都呈现出令人印象深刻的效果。其中, Adaattn 和 SANet 都执行图像特征均值和方差对齐, 专注于生成夸张的艺术效果。但由于没有充分利用 VGG 网络提取的低层特征, 因此产生了较明显的内容信息损失。StyTr2 采用了 Transformer 结构, 并引入 CAPE 作为位置编码, 把目标图像的风格分布合理地映射至内容图像。然而, 由于未引入更多的注意力机制, 其内容保存度有所欠缺。S2WAT 使用条带窗口自注意机制, 消除生成图像中带有网格状的条纹。EMR-RCPN 中的 RCPB 相比 S2WAT 引入 LSA, 在每个子窗口计算自注意, 捕获细粒度和短距离信息, 并与全局下采样自注意交替使用, 增强模型的感受野。

本文对 Twins-SVT+decoder 的结构与 EMR-RCPN 生成的图像进行了对比以及可视化分析, 如图 6 所

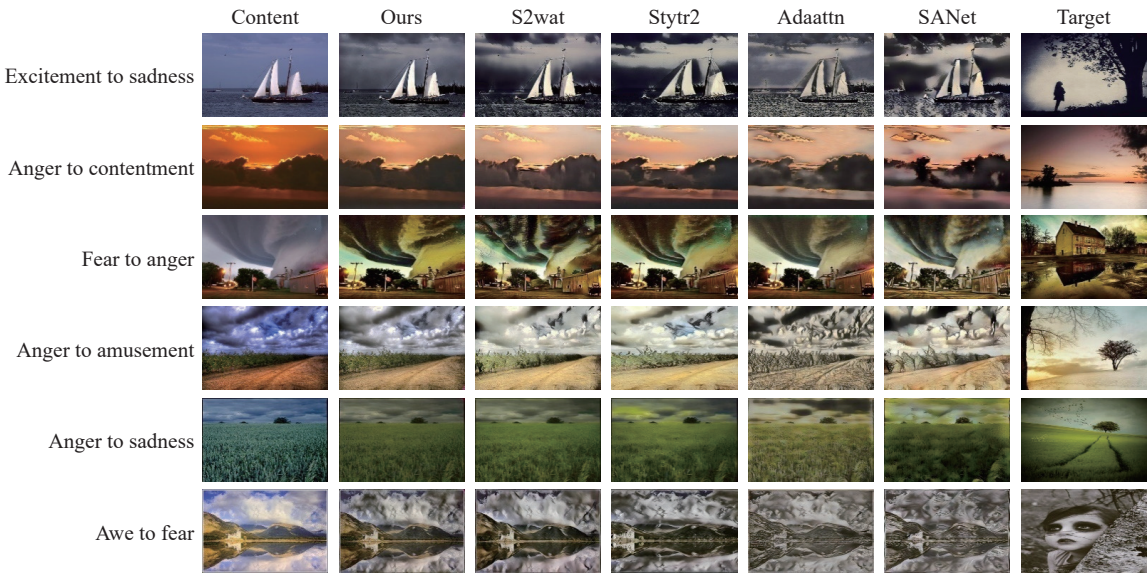


图 5 不同方法生成结果的可视化对比

Fig. 5 Visual comparison of results generated by different methods

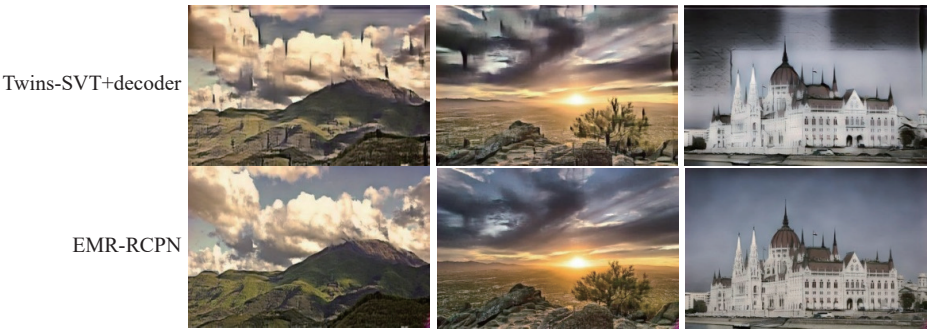


图 6 关于区域上下文感知块的消融实验可视化示例

Fig. 6 Visual example of ablation experiment on regional context perception block

示。由于对比方法中的 LSA 未能聚合足够的上下文信息,导致生成图像出现许多不真实的条纹,它也没有采取如 EMR-RCPN 使用残差连接结构的方式保留图像内容,导致其有较明显的细节丢失。相比之下,EMR-RCPN 编码器中的区域上下文感知块使用多种自注意机制对局部区域的信息进行提取,并使用交叉注意融合不同区域的上下文特征,以消除不合理的条纹。

3 结束语

本文将人类对情感的认知和心理学情感轮作为外部情感知识引入图像情感迁移,提出情感建模关系引导的区域上下文感知增强网络。为了获取丰富的图像上下文信息并有效保留图像的内容结构,本文提出区域上下文感知块用于增强 Twins-SVT 编码器。该模块通过多种自注意机制提取图像不同范围的上下文特征,并通过交叉注意自适应地将其融合,以此获得丰富的图像上下文信息。同时,编码器通过残差连接结构恢复深度特征融合中损失的细节信息,实现更高的内容保留度。此外,所提出的 ETCM 提供了更准确有效的定量评估。在未来的工作中,探究抽象情感与图像特征的关联性,将成为图像情感迁移研究中一个充满希望的方向。

参考文献:

- [1] DENG Y, TANG F, DONG W, *et al.* Stytr2: Image style transfer with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA: IEEE, 2022: 11326-11336.
- [2] LIU S, LIN T, HE D, *et al.* Adaattn: Revisit attention mechanism in arbitrary neural style transfer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021: 6649-6658.
- [3] PARK D Y, LEE K H. Arbitrary style transfer with style-attentional networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA: IEEE, 2019: 5880-5888.
- [4] YANG C K, PENG L K. Automatic mood-transferring between color images[J]. *IEEE Computer Graphics and Applications*, 2008, 28(2): 52-61.
- [5] PENG K C, CHEN T, SADOVNIK A, *et al.* A mixed bag of emotions: Model, predict, and transfer emotion distributions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: IEEE, 2015: 860-868.
- [6] POULI T, REINHARD E. Progressive histogram reshaping for creative color transfer and tone reproduction[C]//Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering. Annecy, France: ACM, 2010: 81-90.
- [7] CHEN T, XIONG W, ZHENG H, *et al.* Image sentiment transfer[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, WA: ACM, 2020: 4407-4415.
- [8] AN J, CHEN T, ZHANG S, *et al.* Global image sentiment transfer[C]//2020 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE, 2021: 6267-6274.
- [9] ZHU S, QING C, CHEN C, *et al.* Emotional generative adversarial network for image emotion transfer[J]. *Expert Systems with Applications*, 2023, 216: 119485.
- [10] ZHANG J, LIU J, DING W, *et al.* Object aroused emotion analysis network for image sentiment analysis[J]. *Knowledge-Based Systems*, 2024, 286: 111429.
- [11] ZHANG J, LIU X, WANG Z, *et al.* Graph-based object semantic refinement for visual emotion recognition[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, 32(5): 3036-3049.
- [12] ZHANG J, LIU X, CHEN M, *et al.* Image sentiment classification via multi-level sentiment region correlation analysis[J]. *Neurocomputing*, 2022, 469: 221-233.
- [13] KIM H R, KANG H, LEE I K. Image recoloring with valence - arousal emotion model[J]. *Computer Graphics Forum*, 2016, 35(7): 209-216.
- [14] LIU D, JIANG Y, PEI M, *et al.* Emotional image color transfer via deep learning[J]. *Pattern Recognition Letters*, 2018, 110: 16-22.
- [15] ZHU S, QING C, XU X. Text-guided generative adversarial network for image emotion transfer[C]//International Conference on Intelligent Computing. Singapore: Springer Nature Singapore, 2023: 506-522.
- [16] DOLAN R J. Emotion, cognition, and behavior[J]. *Science*, 2002, 298(5596): 1191-1194.
- [17] GROSS J J. Emotion regulation: Affective, cognitive, and social consequences[J]. *Psychophysiology*, 2002, 39(3): 281-291.
- [18] SKOTTNIK L, LINDEN D E J. Mental imagery and brain regulation: New links between psychotherapy and neuroscience[J]. *Frontiers in Psychiatry*, 2019, 10: 482129.
- [19] WICKEN M, KEOGH R, PEARSON J. The critical role of mental imagery in human emotion: Insights from fear-based imagery and aphantasia[J]. *Proceedings of the Royal Society B*, 2021, 288(1946): 20210267.
- [20] BARGH J A, CHAIKEN S, GOVENDER R, *et al.* The generality of the automatic attitude activation effect[J]. *Journal of Personality and Social Psychology*, 1992, 62(6): 1118-1124.

- 893.
- [21] MIKELS J A, FREDRICKSON B L, LARKIN G R, *et al.* Emotional category data on images from the international affective picture system[J]. *Behavior Research Methods*, 2005, 37: 626-630.
- [22] KAREN S. Very deep convolutional networks for large-scale image recognition[EB/OL].(2014-9-4)[2015-4-10]. <https://doi.org/10.48550/arXiv.1409.1556>.
- [23] AN J, HUANG S, SONG Y, *et al.* Artflow: Unbiased image style transfer via reversible neural flows[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE, 2021: 862-871.
- [24] 费承, 罗健旭. 基于密集多尺度特征和双注意力模块的皮肤病变分割 [J]. 华东理工大学学报 (自然科学版), 2024, 50(1): 97-105.
- [25] ZHANG C, XU X, WANG L, *et al.* S2WAT: Image style transfer via hierarchical vision transformer using strips window attention[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI Press, 2024, 38(7): 7024-7032.
- [26] 高佳希, 黄海燕. 基于 TF-IDF 和多头注意力 Transformer 模型的文本情感分析 [J]. 华东理工大学学报 (自然科学版), 2024, 50(1): 129-136.
- [27] CHU X, TIAN Z, WANG Y, *et al.* Twins: Revisiting the design of spatial attention in vision transformers[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 9355-9366.
- [28] LIU Z, LIN Y, CAO Y, *et al.* Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021: 10012-10022.
- [29] ZHANG C, DAI Z, CAO P, *et al.* Edge enhanced image style transfer via transformers[C]//Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. Thessaloniki, Greece: ACM, 2023: 105-114.
- [30] ZHAO S, YAO H, GAO Y, *et al.* Predicting personalized emotion perceptions of social images[C]//Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, Netherlands: ACM, 2016: 1385-1394.
- [31] PARK S, KIM J, YE S, *et al.* Dimensional emotion detection from categorical emotion[EB/OL]. (2019-11-6) [2021-9-10]. <https://doi.org/10.48550/arXiv.1911.02499>.
- [32] YOU Q, LUO J, JIN H, *et al.* Building a large scale dataset for image emotion recognition: The fine print and the benchmark[C]//Proceedings of the AAAI Conference On Artificial Intelligence. Phoenix, Arizona, USA: AAAI Press, 2016: 308-314.
- [33] YANG J, SUN M, SUN X. Learning visual sentiment distributions via augmented conditional probability neural network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, California, USA: AAAI Press, 2017: 224-230.
- [34] MACHAJDIK J, HANBURY A. Affective image classification using features inspired by psychology and art theory[C]//Proceedings of the 18th ACM international conference on Multimedia. Firenze, Italy: ACM, 2010: 83-92.
- [35] EKMAN P. An argument for basic emotions[J]. *Cognition & Emotion*, 1992, 6(3/4): 169-200.
- [36] YANG J, LI J, WANG X, *et al.* Stimuli-aware visual emotion analysis[J]. *IEEE Transactions on Image Processing*, 2021, 30: 7432-7445.
- [37] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al.* An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2020-10-22)[2021-6-3]. <https://doi.org/10.48550/arXiv.2010.11929>.
- [38] ZHANG R, ISOLA P, EFROS A A, *et al.* The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 586-595.
- [39] WANG Z, BOVIK A C. A universal image quality index[J]. *IEEE Signal Processing Letters*, 2002, 9(3): 81-84.
- [40] WANG Z, BOVIK A C, SHEIKH H R, *et al.* Image quality assessment: from error visibility to structural similarity[J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 600-612.
- [41] HEUSEL M, RAMSAUER H, UNTERTHINER T, *et al.* Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. *Advances in Neural Information Processing Systems*, 2017, 30: 1-12.

Regional Context Perception Enhanced Network for Image Emotion Transfer

ZHU Jixiang, ZHANG Jing

(School of Information Science and Engineering, East China University of Science and Technology,
Shanghai 200237, China)

Abstract: To achieve accurate and effective image emotion transfer, a region context aware enhancement network guided by emotion modeling relationships is proposed. Under the guidance of external emotion knowledge, multiple losses are combined to achieve accurate image induced emotion transfer. In this network, a novel regional context perception block (RCPB) is introduced to enhance the Twins-SVT encoder. It extracts context features of different receptive fields in images through various Locally-grouped Self Attention, Axis-wise Self Attention, and Neighbor Self Attention, and adaptively fuses them through cross attention to comprehensively integrate image information. On this basis, the lost information in the fusion of deep features are restored through residual connections to more accurately preserve the image content. Additionally, a novel emotion-wheel guided module (EGM) is proposed, which uses the emotion distribution in the emotion wheel to guide the model in accurately transferring image emotions. To accurately and effectively evaluate the ability of the model to transfer image emotion, an innovative emotion transfer comprehensive metric (ETCM) is proposed, which evaluates the effect of image emotion transfer from multiple perspectives, including emotion categories, emotion polarities, and the position of emotions in the emotion wheel. To effectively evaluate the model effectiveness of the model, a new emotion transfer dataset FATE is constructed based on four widely used emotion datasets with different styles: FI, Twitter-LDL, Emotion6, and Artphoto. The extensive experiments on FATE have fully validated the effectiveness of the proposed method and demonstrated its superiority over other compared methods.

Key words: image emotion transfer; regional context perception enhanced; emotion modeling relationship; emotion transfer comprehensive metrics; self attention mechanism

(责任编辑: 李娟)