

中图法分类号: TP37 文献标识码: A 文章编号: 1006-8961(2023)11-3536-14

论文引用格式: Sun M T, Dai L Q and Tang J H. 2023. Transformer-based multi-style information transfer in image processing. Journal of Image and Graphics, 28(11):3536-3549(孙梅婷, 代龙泉, 唐金辉. 2023. 基于Transformer方法的任意风格迁移策略. 中国图象图形学报, 28(11):3536-3549)[DOI:10.11834/jig.211237]

基于Transformer方法的任意风格迁移策略

孙梅婷, 代龙泉*, 唐金辉

南京理工大学计算机科学与工程学院, 南京 210094

摘要: 目的 任意风格迁移是图像处理任务的重要分支, 卷积神经网络作为其常用的网络架构, 能够协助内容和风格信息的提取与分离, 但是受限于卷积操作感受野, 只能捕获图像局部关联先验知识; 而自然语言处理领域的Transformer网络能够突破距离限制, 捕获长距离依赖关系, 更好地建模全局信息, 但是因为需要学习所有元素间的关联性, 其表达能力的提高也带来了计算成本的增加。鉴于风格迁移过程与句子翻译过程的相似性, 提出了一种混合网络模型, 综合利用卷积神经网络和Transformer网络的优点并抑制其不足。**方法** 首先使用卷积神经网络提取图像高级特征, 同时降低图像尺寸。随后将提取的特征送入Transformer中, 求取内容特征与风格特征间的关联性, 并将内容特征替换为风格特征的加权和, 实现风格转换。最后使用卷积神经网络将处理好的特征映射回图像域, 生成艺术化图像。**结果** 与5种先进的任意风格迁移方法进行定性和定量比较。在定性方面, 进行用户调查, 比较各方法生成图像的风格化效果, 结果表明本文网络生成的风格化图像渲染效果更受用户喜爱; 在定量方面, 比较各方法的风格化处理速度, 结果表明本文网络风格化速率排名第3, 属于可接受范围内。此外, 本文与现有的基于Transformer的任意风格迁移方法进行比较, 突出二者间差异; 对判别网络进行消融实验, 表明判别网络的引入能够有效提升图像的光滑度和整洁度; 最后, 将本文网络应用于多种风格迁移任务, 表明本文网络具有灵活性。**结论** 本文提出的混合网络模型, 综合了卷积神经网络和Transformer网络的优点, 同时引入了判别网络, 使生成的风格化图像更加真实和生动。

关键词: 计算机视觉; 图像处理; 任意风格迁移; 注意力机制; Transformer

Transformer-based multi-style information transfer in image processing

Sun Meiting, Dai Longquan*, Tang Jinhui

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Abstract: Objective The multi-style transfer technique can be applied and focused on the stylized image via visual styles-transferred content image. The stylized image can be used to preserve original content structure and it has the similar features with the style image. Style transfer can be as one of the essential branches of image processing. Conventional style transfer methods are oriented for style rendering in terms of low-quality information processing. Current deep learning based convolutional neural network (CNN) has been adopted in the style transfer domain. To balance and re-integrate images' content and style information, the CNN can be used to extract content features and style features. However, due to the constrained range of visual perception of the convolutional layers, it can capture local associations only. To model its

收稿日期: 2022-01-10; 修回日期: 2022-08-27; 预印本日期: 2022-09-03

* 通信作者: 代龙泉 dailongquan@njut.edu.cn

基金项目: 国家自然科学基金项目(62072238, 61732007, 62102181); 中央高校基本科研业务费专项资金资助(30922010911)

Supported by: National Natural Science Foundation of China (62072238, 61732007, 62102181); Fundamental Research Funds for the Central Universities (30922010911)

global information, the Transformer network proposed in natural language processing (NLP) can capture long-distance dependencies. However, its expression-related requirement is still challenged for computational cost because of correlation-learned between all input elements. Furthermore, the Transformer has its slower convergence performance due to a lack of image prior. Given the similarity between the image style transfer process and the sentence translation process, we develop a CNNs-based hybrid network in terms of Transformer network. **Method** The network we proposed consists of four aspects in relevance to: encoding, style transformation, decoding, and the discriminative network. For network-encoded: convolutional layers are used to extract the high-level image features through lowering the image size. To exact more value of the pixels, features-extracted are more sensitive to semantic and consistent information in related to such specific objects and content in the image. However, the lower quality information of the image, such as lines and textures, is beneficial to reflect the stylistic features. Thus, the network-encoded can be added into the residual connection to enrich feature representation. For Transformer network structure-built style transformation network: it consists of three subparts: content encoder, style encoder, and the decoder. The content and style encoder is paid attention to global information-added for each of content features and style features. The decoder can be used to optimize the original content features in related to the weighted sum of style features and stylized features can be generated further. For network-decoded: to up-sample stylized features back to the original image size and generate a final stylized image, topology-interpolated operation is interlinked to this sort of symmetric structure with the encoding network. For discriminative network: it is focused on distinguishing between the generated and natural style images. **Result** To verify the style transfer evaluation criteria, qualitative and quantitative comparative analysis is carried out with several other related flexible style transfer methods. The qualitative analysis consists of two parts: performance comparison and user experience. The performance comparison can show that the proposed network can generate more smooth and clear stylized images. The user-oriented result can show its user-preference ability as well. The results of the stylizing speed comparison in the quantitative comparison can demonstrate that the speed of the proposed network ranks is in the acceptable threshold. Additionally, the speed can keep its stability as the image size grows from 256 pixels to 512 pixels. An ablation experiment is designed to verify the effectiveness of the discriminative network as well. Its results show that the discriminative network-introduced can yield the network to get extraction ability better and generate more realistic images. To demonstrate the flexibility of the network we proposed, the trained network can be used for other related style transfer tasks, including content-style tradeoff, style interpolation, and region painting. **Conclusion** A hybrid network is facilitated and mutual-benefits are shown between CNNs and Transformer network. Experiment results show that the network we proposed can optimize the transferring speed further. The stylized images have its potentials for smaller image sizes (e. g., 256 pixels) in terms of its content structure and stylistic features.

Key words: computer vision; image processing; multi-style image information transfer; attention mechanism; Transformer

0 引言

任意风格迁移技术(arbitrary style transfer)旨在寻求一种通用的风格转换方法,对于给定的任意一幅风格图像,都能依据其样式特点对内容图像进行渲染,生成同时具备内容图像语义结构和风格图像样式纹理的艺术化合成图像。风格迁移作为图像处理领域的重要分支,广泛应用于日常生活中的照片和视频处理软件中,为用户提供多种多样的视觉效果,而任意风格迁移方向提高了风格迁移技术的通用性和灵活性,因此受到广泛关注。

在风格迁移研究中,物体、结构和语义被广泛视为图像的内容表示,颜色、纹理和笔触被广泛视为图像的风格表示。因此,传统风格迁移技术(Haeberli, 1990; Hertzmann, 1998; Efros 和 Freeman, 2001; Hertzmann 等, 2001; Winnemöller 等, 2006)通过对纹理、线条等图像低级信息进行抽象处理来实现图像的艺术化转变。虽然有一定的风格化成果,但是因为只利用了图像的表征信息,缺乏整体语义的思考,因此存在视觉效果不佳、处理时间长等不足。

随着深度学习在各个领域开花结果,神经风格迁移(neural style transfer)方法应运而生,利用卷积神经网络(convolutional neural network, CNN)自适应

提取图像特征,能够更好地建模图像的内容和风格表征。Gatys 等人(2016)首次将卷积神经网络应用到风格迁移任务中,将网络提取的图像特征视为内容表示,将特征间的相关性视为风格表示,通过分离与重组,融合内容图像和风格图像,虽然能够产生良好的风格效果,但是其中用到的图像迭代优化过程存在风格化速度慢、耗时长等弊端。考虑到生成对抗网络(generative adversarial networks, GAN)(Goodfellow 等, 2014)优秀的图像生成能力,一些工作(Zhu 等, 2017; Chen 等, 2018; Hicsonmez 等, 2020; Koto-venko 等, 2019)将 GAN 应用于风格迁移领域,通过训练网络模型,实现快速风格化。尽管这类方法提高了风格化速率,同时能够生成栩栩如生的风格化图像,但其网络模型受限于 GAN 网络框架,只能实现有限风格种类的渲染效果,缺乏灵活性,如果要应用于不同风格则需要对每一种风格单独训练一个模型,增加了训练成本。得益于条件生成网络(conditional adversarial network, cGAN)(Mirza 和 Osindero, 2014)的出现, Yanai(2017)使用条件网络直接从风格图像中生成条件信号,控制网络生成的风格,实现了任意风格的转换。

GAN 中的判别器学习分辨特定类别图像的真实性,在一定程度上约束了生成图像风格的多样性。为了提高生成网络的灵活性,实现任意的艺术化风格转换效果,后续工作更加关注图像的编码—转换—解码过程,旨在寻求能够更好地分离、融合内容信息与风格信息的方法。任意风格迁移方法中,在内容图像中引入风格信息的主要手段可以大致分为基于统计信息和基于图像块两种。前者使用图像分布统计数值表示风格信息(Huang 和 Belongie, 2017; Li 等, 2017; Luan 等, 2017; Wang 等, 2020; Xie 等, 2020),如均值、方差和协方差等,通过整体调整内容图像的分布,使其符合风格图像数据分布,从而获得与风格图像相同的样式特征。后者将图像划分为图像块,用图像块作为风格特征的表示(Chen 和 Schmidt, 2016; Sheng 等, 2018),通过将风格图像块替换到内容图像中,使得内容图像具备风格图像的纹理表征。

CNN 作为风格迁移领域中使用的主流网络框架,能够有效提取图像的内容特征和风格特征,但是卷积操作限制了网络的感受野范围,不利于图像全局信息的捕获,若需要获取更大范围的感受野,则需

要堆叠更深的网络层,不利于网络训练。而 Transformer(Vaswani 等, 2017)恰好能弥补卷积网络的不足。Transformer 网络结构最早在自然语言处理领域(natural language processing, NLP)的机器翻译任务中提出,它的网络主体结构由注意力网络层堆叠而成,这使得它可以轻松捕获句子中单词间存在的长距离依赖关系,加深网络对每个单词的理解,提高句子翻译的合理性和准确性。通过抽象句子翻译过程,可以发现基于图像块的风格迁移方法与句子翻译任务有异曲同工之处。句子翻译过程可以看做是单词到单词的转换过程,而基于图像块的风格迁移过程可以视为图像块到图像块(特征到特征)的替换问题,如图 1 所示。句子可以表示为由单词构成的序列,则句子翻译过程即为依据输入的单词序列信息预测输出单词的过程;依此思路,将图像按固定尺寸划分,则图像可以表示成由图像块构成的序列,每次从风格图像块序列中选择合适的风格图像块替换内容图像序列中对应的内容图像块,将内容图像块序列中所有的图像块都替换完毕后得到风格化图像块序列,将序列重新排列后获得最终的风格化图像。

依据上述分析,将 Transformer 应用于风格迁移任务是合理的和可行的。但是,直接使用 Transformer 完成全部的图像处理任务,仍存在一些问题。1)Transformer 缺乏归纳偏置(inductive bias),难以获取图像先验信息,因此在训练中收敛速度较慢;2)缺乏有效的特征提取手段;3)图像相较于句子,序列长度更长,导致注意力网络层计算消耗更大。而卷积网络能够弥补 Transformer 存在的不足,因此将二者结合能更好地实现风格迁移效果。综上所述,本文主要贡献包括以下几点:1)提出一个用于任意风格迁移任务的混合网络,同时结合 CNN 和 Transformer 网络优势,使用卷积操作提取图像特征,提供图像先验知识,便于局部信息的融合;使用 Transformer 捕获特征间的内在联系,丰富特征表示,提供全局信息交互能力;2)引入判别网络,更好地度量生成的风格化图像和现实中真实风格图像间的差异,减少不合理纹理的产生,提升画面渲染的光滑度;3)通过与其他任意风格迁移方法进行定性和定量比较,表明本文网络能够生成画面更光滑、真实性更高、风格特征更明显的高质量风格化图像。

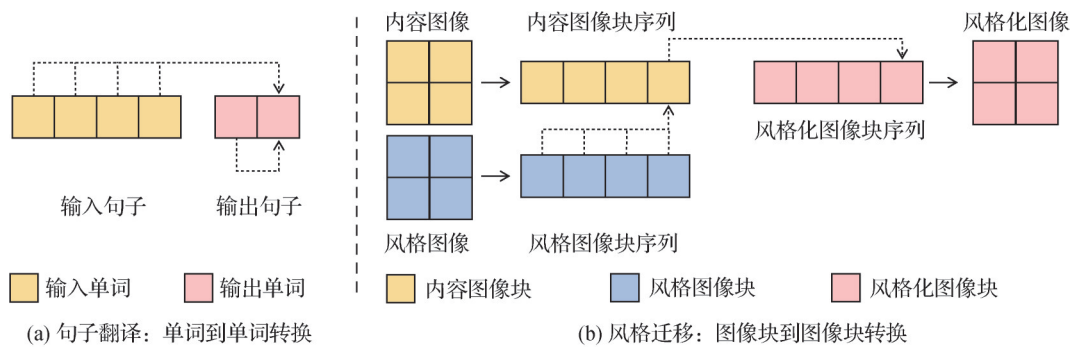


图1 句子翻译和基于图像块的风格迁移之间的相似性

Fig. 1 Similarity between sentence translation and patch-based style transfer

((a) sentence translation: word-to-word translation; (b) style transfer: patch-to-patch translation)

1 相关工作

鉴于CNN能够建模复杂的、手工设计难以捕获和表示的图像特征,已经广泛用做图像处理任务中的编码和解码网络。图像输入编码网络中提取特征,特征经由转换网络生成风格化特征,风格化特征再输入到解码网络中得到最终的风格化图像,这种编码—转换—解码的处理流程是任意风格迁移任务中常使用的网络框架。

在结构丰富多样的CNN中,预训练的VGG (Visual Geometry Group) 网络 (Simonyan 和 Zisserman, 2015) 是任意风格迁移方法中常见的编码网络结构。它作为特征提取器,为网络提供可靠的图像特征提取能力,同时因为其本身已经具备良好的学习能力,所以在网络训练过程中,它可以指导网络其他部分的学习,稳定网络的训练过程,加速网络模型的收敛。不同于直接引入VGG网络的方法 (Huang 和 Belongie, 2017; Li 等, 2017; Luan 等, 2018; Park 和 Lee, 2019; Yao 等, 2019; Wang 等, 2020; Chen 等, 2021a; Deng 等, 2021), Sheng 等人 (2018) 综合考虑网络编码与解码过程,在VGG网络结构中增添中间链接结构,将提取的特征信息引入到图像解码过程中,增强了解码网络的稳定性与图像重建能力。Chen 和 Schmidt (2016) 分别在图像空间与特征空间进行风格迁移实验,表明直接对图像进行处理生成的风格化结果,虽然有颜色的改变,但是风格特征并不明显;而在特征空间进行处理生成的风格化结果具备更加优越的风格特征,说明在任意风格迁移任务中,使用CNN表示图像复杂信息是必要的。

除了CNN,本文还使用了其他理论知识和关键技术。

1.1 判别网络

网络训练时,由于人力与物力的局限,有时难以收集明确的标签数据集,Goodfellow 等人 (2014) 提出GAN网络,在其中使用判别网络分辨生成结果的真实性,以此促使生成网络提升图像生成能力。

在有限风格种类迁移中,Chen 等人 (2018) 提出了专门用于图像卡通化的网络结构,促使生成图像具备卡通图像高度简化、边缘清晰和色彩平滑的特点。Kotovenko 等人 (2019) 通过在内容图像与风格样本中寻找相似语义内容,使网络学习特定风格是如何改变内容细节的,从而提高网络对其他内容对象的渲染质量。Hicsonmez 等人 (2020) 则将图像的形态特征和边缘轮廓等低级信息引入风格转换过程中,协助网络更好地保留内容,实现插图类型图像的生成。

在任意风格种类迁移中,Chen 等人 (2021a) 使用判别网络分辨真实风格图像和虚假的生成图像,提高了风格化图像中颜色渲染的和谐性与纹理样式的合理性。

1.2 注意力机制

注意力机制 (attention mechanism) 即对观察事物的不同部分倾注不同程度的关注,依此方式过滤冗余信息,保留关键信息。在人类日常生活中,注意力协助大脑快速、高效地处理海量信息,加快反应速度,辅助关键信息记忆。为了使神经网络学习数据间存在的关联性,提高数据处理能力,有些工作引入了注意力机制。

在计算机视觉领域, Lee 等人 (2019) 应用注意

力调整风格权重,使网络关注重要的风格特征,提高卷积神经网络的表示能力。Fu 等人(2019)利用注意力建模全局信息的能力,同时捕获特征在空间和通道维度的上下文关系,生成了更精确的图像分割结果。Zhang 等人(2019)将自注意力融入 GAN 网络,与卷积操作相辅相成,实现长距离、多尺度建模。

在风格迁移领域,注意力机制能够帮助网络关注内容图像的主要语义结构和风格图像的关键纹理笔触。Park 和 Lee(2019)提出风格注意力网络,将内容特征表示为风格特征的加权和,灵活且有效地融合了风格信息。为了丰富生成图像中风格笔触的多样性,Yao 等人(2019)使用 K-means 方法将内容特征聚类分区,并用注意力捕获分区间依赖关系,指导不同尺度特征的融合过程。Deng 等人(2020)在内容图像上使用空间注意力增强内容表征,在风格图像上使用通道注意力增强风格表征,自适应地分离内容和风格特征,从而生成语义化渲染效果。

注意力机制协助网络使用主要风格特征绘制主要内容结构,实现语义感知的风格化渲染效果。

1.3 Transformer

图像可以看做由像素点构成的“句子”,相较于卷积操作需要堆叠多层卷积层来获得更大感受野内像素间的依赖关系,Transformer 能够更方便地建模长距离依赖,因此越来越多的研究人员探索将 Transformer 应用在图像处理任务中(刘花成 等,2022)。

Transformer 遵循编码—解码的网络结构,整体结构如图 2 所示。网络由多头注意力网络层(multi-head self-attention, MSA)、前向网络层(feed-forward network, FFN)和层标准化网络层(layer normalization, LN)构成。图 2 中的 l 符号表示编码层和解码层的层数,Add 符号表示相加操作,Embedding 网络层将序列映射到更高维度。Vaswani 等人(2017)介绍了 Transformer 网络的具体实现细节。

在分类、分割和检测任务中,Dosovitskiy 等人(2021)将图像划分为固定大小的图像块,构造与句子类似的序列化数据,输入 Transformer 编码器处理后,得到图像分类结果。Liu 等人(2021)以滑动窗口形式计算图像的局部注意力,降低网络计算量。

在图像生成、超分辨率等需要生成图像的任务中,Parmer 等人(2018)将图像生成任务视为像素点回归任务,搭建 Image Transformer 网络,以逐像素预

测的方式生成图像。Esser 等人(2021)将 Transformer 优秀的表达能力与 CNN 自带的归纳偏置相结合,实现高分辨率图像合成结果。上述两种方法均以回归方式生成图像,前者使用像素,后者使用特征,因为需要逐个预测,因此生成速度较慢。

Chen 等人(2021b)使用卷积网络处理图像,使用 Transformer 处理特征,在超分辨率、去雨和去噪等图像处理任务上均取得了优异成果。Jiang 等人(2021)使用 Transformer 网络结构搭建了第一个没有卷积操作的生成对抗网络 TransGAN,使用数据增强、多任务联合训练和局部注意力初始化等策略,增强图像块间连接的光滑度。可以看出,没有卷积网络的辅助,则需要更多的策略辅助网络训练。

尽管 Transformer 在图像处理任务中得到了广泛关注,但是将其应用于任意风格迁移领域的研究屈指可数。鉴于风格迁移过程与句子翻译过程的相似性,引入 Transformer 实现特征转换,有助于为风格迁移领域注入新的活力。

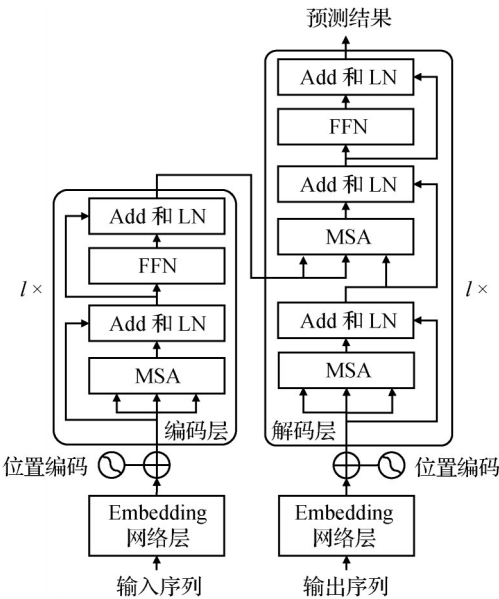


图 2 Transformer 网络结构

Fig. 2 Network architecture of Transformer

2 研究方法

为了挖掘 Transformer 在任意风格迁移领域的潜力,本文提出了融合 Transformer 网络结构的任意风格迁移网络,将从网络结构和损失函数两方面对其进行详细介绍,同时为了便于读者理解本

文方法与Transformer网络结构之间存在的联系, Transformer相关网络层符号表示均与1.3节保持一致。

2.1 网络结构

图3展示了提出的任意风格迁移网络的整体结构,该网络输入内容图像 I_c 和风格图像 I_s ,输出风格化图像 I_{cs} 。网络整体遵循对抗生成网络框架,图中生成网络 G 负责学习生成高质量的风格化图像,判别网络 D 则负责学习判断输入是否为真实风格图像。

生成网络 G 由3部分构成,分别为图像编码网络、风格转换网络和图像解码网络,在图3中分别以Enc、T和Dec表示。图像的风格化过程如下:

首先,将内容图像 I_c 和风格图像 I_s 输入编码网络Enc中,提取内容特征 f_c 和风格特征 f_s ,具体为

$$f_c = \text{Enc}(I_c) \quad (1)$$

$$f_s = \text{Enc}(I_s) \quad (2)$$

随后,将 f_c 和 f_s 输入风格转换网络T中,依据二者元素间的相关性,将内容特征替换为对应的风格特征,得到风格化特征 f_{cs} ,具体为

$$f_{cs} = T(f_c, f_s) \quad (3)$$

最后,风格化特征 f_{cs} 经由解码网络Dec解码回图像域,生成风格化图像 I_{cs} ,具体为

$$I_{cs} = \text{Dec}(f_{cs}) \quad (4)$$

上述生成过程为常见的编码—转换—解码风格迁移方式,为了促使生成网络生成更加真实的艺术化结果,增添一个判别网络,度量生成的风格化图像与真实风格图像间的分布差异,拉近二者的距离。判别网络 D 在训练过程中,学习为真实风格图像分类标签1(real),为生成的风格化图像分类标签0(fake),即

$$D(I_s) \rightarrow 1, \quad D(I_{cs}) \rightarrow 0 \quad (5)$$

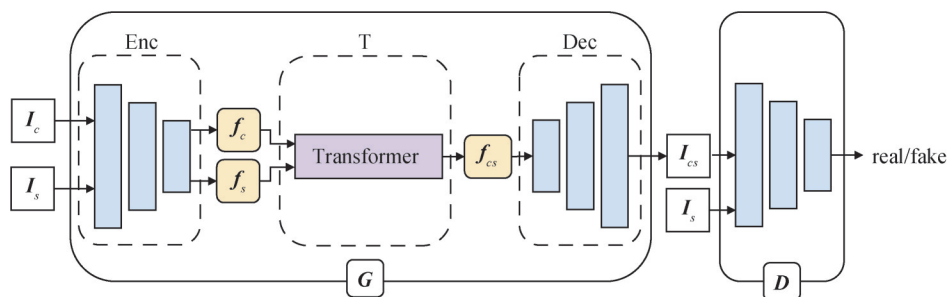


图3 基于Transformer的任意风格迁移网络结构

Fig. 3 Network architecture for Transformer-based arbitrary style transfer

2.1.1 图像编码网络

卷积网络种类众多,鉴于风格迁移任务不仅需要图像深层信息理解内容语义结构,而且需要图像浅层信息辅助风格纹理笔触的表现,因此使用残差网络(He等,2016)作为编码网络,同时将网络中第1层 7×7 大小的卷积核改为 3×3 ,保留更多的线条纹理细节。因为风格转换网络中用到了Transformer网络结构,为了降低网络中注意力网络层的计算量,在残差网络块中使用平均池化层,降低图像尺寸。平均池化层缩放因子设置为2,设训练时输入图像宽高为 2^x 像素,编码出来的特征宽高固定为 2^y 像素,则编码网络中包含的残差块有 $x-y$ 层。

将内容图像 $I_c \in \mathbf{R}^{3 \times H \times W}$ 和风格图像 $I_s \in \mathbf{R}^{3 \times H \times W}$ 输入编码网络中,生成尺寸较小的内容特征 $f_c \in \mathbf{R}^{c \times h \times w}$ 和风格特征 $f_s \in \mathbf{R}^{c \times h \times w}$ 。其中, H 和 W 分

别表示图像的高度和宽度, c, h 和 w 分别表示提取特征的通道数、高度和宽度, $H = W = 2^x, h = w = 2^y$ 。

2.1.2 风格转换网络

原始Transformer网络包含编码器和解码器两部分,编码器建模句子单词间的关联性,解码器部分依据编码器编码结果及之前输出的单词,预测当前位置单词的概率分布。依据解码部分的特点,可以将其视为基于块替换的风格迁移方法,即依据风格特征,替换内容特征。该部分网络尽量遵循原始Transformer结构(Vaswani等,2017),考虑到风格迁移的输入为两幅图像,为了使网络更好地学习内容图像和风格图像的特点,在原始Transformer结构的基础上额外增加一个编码器,构成一个由两个编码器和一个解码器组成的风格转换网络。两个编码器为内容编码器CTE(content Transformer encoder)和

风格编码器 STE (style Transformer encoder), 解码器用 TD (Transformer decoder) 表示, 转换网络结构如图 4 所示。

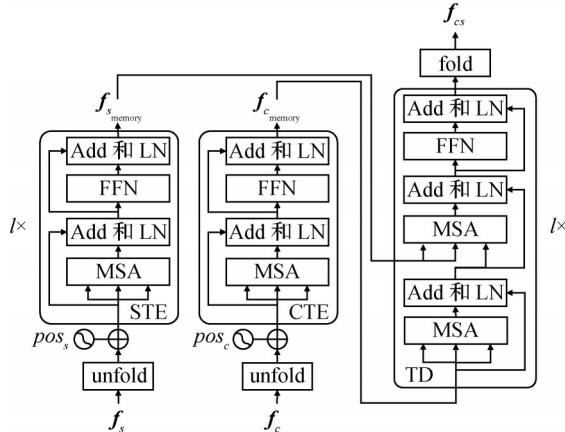


图 4 风格转换网络结构

Fig. 4 The style transformation network structure

风格转换网络接收内容特征 f_c 和风格特征 f_s 作为其输入, 输出风格化特征 $f_{cs} \in \mathbf{R}^{c \times h \times w}$ 。网络中的 unfold 操作使用高度和宽度为 P 的滑动窗口在特征上以步长 P 进行滑动, 滑动过程中将窗口内所有数据合并并在通道维度上, 将输入的特征展开为特征序列的形式。经过 unfold 操作处理后, 得到内容特征序列 $f_{c \rightarrow \text{seq}} \in \mathbf{R}^{c' \times L}$ 和风格特征序列 $f_{s \rightarrow \text{seq}} \in \mathbf{R}^{c' \times L}$, 其中, $c' = c \times P^2$ 为特征块编码长度, $L = hw/P^2$ 为序列长度。

为了利用序列中蕴含的特征顺序信息, 分别为内容特征序列添加内容位置编码 pos_c , 为风格特征序列添加风格位置编码 pos_s , 其中位置编码与序列形状相同, 便于二者相加。添加位置信息后将内容特征序列 $f_{c \rightarrow \text{seq}}$ 输入内容编码器 CTE 中处理, 将风格特征序列 $f_{s \rightarrow \text{seq}}$ 输入风格编码器 STE 中处理, 得到内容编码器的输出 f_{c_memory} 和风格编码器的输出 f_{s_memory} , 上述过程具体表示为

$$f_{c_memory} = CTE(f_{c \rightarrow \text{seq}} + pos_c) \quad (6)$$

$$f_{s_memory} = STE(f_{s \rightarrow \text{seq}} + pos_s) \quad (7)$$

将内容编码器和风格编码器输出的 f_{c_memory} 和 f_{s_memory} 作为解码器 TD 的输入, 求取风格化特征序列 $f_{cs \rightarrow \text{seq}}$, 具体为

$$f_{cs \rightarrow \text{seq}} = TD(f_{c_memory}, f_{s_memory}) \quad (8)$$

依靠解码器中第 2 层注意力网络层 MSA (multi-

head self-attention) 可以达到引入风格信息的目的, 该过程可表示为

$$MSA(Q_c, K_s, V_s) = \text{softmax}\left(\frac{Q_c K_s^T}{\sqrt{c'}}\right) V_s \quad (9)$$

$$Q_{cs} = MSA(Q_c, K_s, V_s) + Q_c \quad (10)$$

式中, Q_c 为内容信息, 来自 f_{c_memory} , K_s 和 V_s 均为风格信息, 来自 f_{s_memory} 。Vaswani 等人 (2017) 具体介绍了 MSA 注意力操作, 该过程可理解为求取内容信息 Q_c 和风格信息 K_s 间的相关性, 使用 softmax 激活函数将二者的相关性转换为权重值, 将权重值与风格信息 V_s 相乘得到应引入的风格信息量, 最后将内容信息 Q_c 加上风格信息量, 得到同时具有内容信息和风格信息的 Q_{cs} 。 Q_{cs} 为 $f_{cs \rightarrow \text{seq}}$ 中的信息。

使用合并操作 fold 将风格化特征序列 $f_{cs \rightarrow \text{seq}}$ 调整回图像的排列形式, 得到风格化特征 $f_{cs} \in \mathbf{R}^{c \times h \times w}$, 其形状与输入的内容特征相同, 便于后续图像解码网络结构的设计与实现。

2.1.3 图像解码网络

解码网络同样由残差网络块搭建而成, 与编码网络相反, 解码网络需要将尺寸较小的特征上采样到原始图像的大小, 因此在该部分残差块中使用最近邻插值操作成倍增加特征尺寸。解码网络结构与编码网络结构对称, 便于网络学习对称的映射关系。

将风格转换网络输出的风格化特征 f_{cs} 输入解码网络中, 生成风格化图像 I_{cs} 。其中, $I_{cs} \in \mathbf{R}^{3 \times H \times W}$ 与原始内容图像 I_c 尺寸相同。

2.1.4 判别网络

通常, 任意风格迁移任务只由生成网络组成, 使用风格损失函数约束生成的风格化效果。然而在实际中, 使用数据分布表示风格信息的方式, 对风格特点进行抽象和量化, 可能会损失一些分布无法建模的风格特征信息。为了更好地度量生成的风格化图像和现实中真实风格图像间的差异, 本文引入一个判别网络, 学习分辨真实和虚假的风格图像, 从而督促生成网络生成更加栩栩如生、贴近现实的艺术化图像。

判别网络遵循 Chen 等人 (2021a) 工作中的设置, 由 3 个判别器组成, 分别分辨原始图像尺寸、1/2 图像尺寸和 1/4 图像尺寸下输入图像的真实性。通过综合不同视野范围下的判别结果, 强化网络的判别能力, 实现多尺度风格差异度量。

2.2 损失函数

风格迁移的目的是在保证内容图像结构轮廓的前提下,使其具有与风格图像相同的色调和笔触,因此,使用内容损失函数和风格损失函数对生成结果进行约束。

使用感知内容损失(Johnson等,2016)约束生成图像中的结构保留效果,该损失在高维空间中度量特征间的差异,能够在保留语义信息的情况下,为结构轮廓提供合理的风格形变,其定义为

$$L_c = \sum_{i=4,5} \|\bar{\phi}_i(I_c) - \bar{\phi}_i(I_s)\|_2 \quad (11)$$

式中, $\phi_i(\cdot)$ 表示预训练 VGG-19 网络(Simonyan 和 Zisserman, 2015)中 Relui-1 层提取的特征, $\bar{\phi}_i(\cdot)$ 表示特征的均值一方差标准化版本,使用标准化版本是为了降低风格信息的影响。

风格损失定义为

$$L_s = \sum_{i=1}^N (\|\mu(\phi_i(I_s)) - \mu(\phi_i(I_{cs}))\|_2 + \|\sigma(\phi_i(I_s)) - \sigma(\phi_i(I_{cs}))\|_2) \quad (12)$$

式中, $\mu(\cdot)$ 和 $\sigma(\cdot)$ 表示特征的均值和方差, N 为使用的网络层数量,在本文中设置为 5,即在多尺度特征上约束生成图像的风格分布。

任意风格迁移过程是无监督网络训练过程,因为缺少明确真值图像的强力约束,所以只能将风格信息建模为均值一方差的分布来约束风格化效果。这种风格损失考虑了图像整体风格分布情况,但是缺乏对局部纹理细节、笔触的考量。因此,本文引入一个判别网络,协助分辨人眼观察不到的风格特征和风格分布难以表示的细节。判别网络带来的对抗损失函数定义为

$$L_{adv} = E_{I_c \sim S} [\log(D(I_s))] + E_{I_c \sim C, I_s \sim S} [\log(1 - D(G(I_c, I_s)))] \quad (13)$$

式中, $G(I_c, I_s) = Dec(T(Enc(I_c), Enc(I_s)))$, 表示生成图像, D 为判别器,判别输入的图像是真实的还是网络生成的, E 为期望, C 和 S 分别表示内容图像和风格图像的集合,有 $I_c \in C, I_s \in S$ 。判别网络通过与生成网络进行对抗训练,学习更能分辨风格信息的关键图像特征;生成网络为了迷惑判别网络,学习生成更接近真实图像的艺术化结果。

网络的总损失函数由内容损失、风格损失和对

抗损失组成,即

$$L_{total} = \lambda_c L_c + \lambda_s L_s + \lambda_{adv} L_{adv} \quad (14)$$

式中, λ_c , λ_s 和 λ_{adv} 分别表示内容损失权重、风格损失权重和对抗损失权重。

3 实验与分析

3.1 实验设置

本文网络基于 PyTorch 深度学习框架搭建,使用的內容数据集为 MS-COCO (Microsoft common objects in context), 风格数据集为 WikiArt (WikiArt dataset), 每个数据集均包含大约 80 000 幅训练图像。在网络训练阶段,输入图像统一裁剪为 256 × 256 像素。网络学习率设置为 0.000 1,使用两个 Adam 优化器分别对生成网络和判别网络进行优化,前者参数设置 $\beta_1 = 0.9$ 和 $\beta_2 = 0.999$,后者参数设置为 $\beta_1 = 0.5$ 和 $\beta_2 = 0.999$ 。风格转换网络中,网络层数 l 设置为 6。总损失函数中,权重参数 λ_c , λ_s 和 λ_{adv} 均设置为 1。

3.2 实验结果

从定性和定量两方面对提出的任意风格迁移方法进行分析与比较,并进行消融实验,验证引入判别网络的必要性。

3.2.1 风格化效果比较

与 5 个先进的任意风格迁移方法进行比较,包括 AdaIN (adaptive instance normalization) (Huang 和 Belongie, 2017)、WCT (whitening and coloring transforms) (Li 等, 2017)、SANET (style-attentional network) (Park 和 Lee, 2019)、AAMS (attention-aware multi-stroke) (Yao 等, 2019) 和 A-Net (Avatar-Net) (Sheng 等, 2018),图 5 展示了各方法生成的风格化图像。

图 5(a)和图 5(b)分别为内容图像和风格图像,剩余依次为各任意风格迁移方法生成的结果。画家作品拥有繁复多样的色彩、笔触等纹理信息,AdaIN 整体调整内容特征分布实现风格渲染,能够保留较好的结构轮廓,但是因为简化了风格信息的表示,所以生成结果风格特点不够明显,如图 5(c)所示。WCT 通过修改特征 2 维统计信息实现风格化效果,能更好地捕获风格特点,但是在高维空间调整风格,图像的内容特征也更容易受到影响,如图 5(d)所

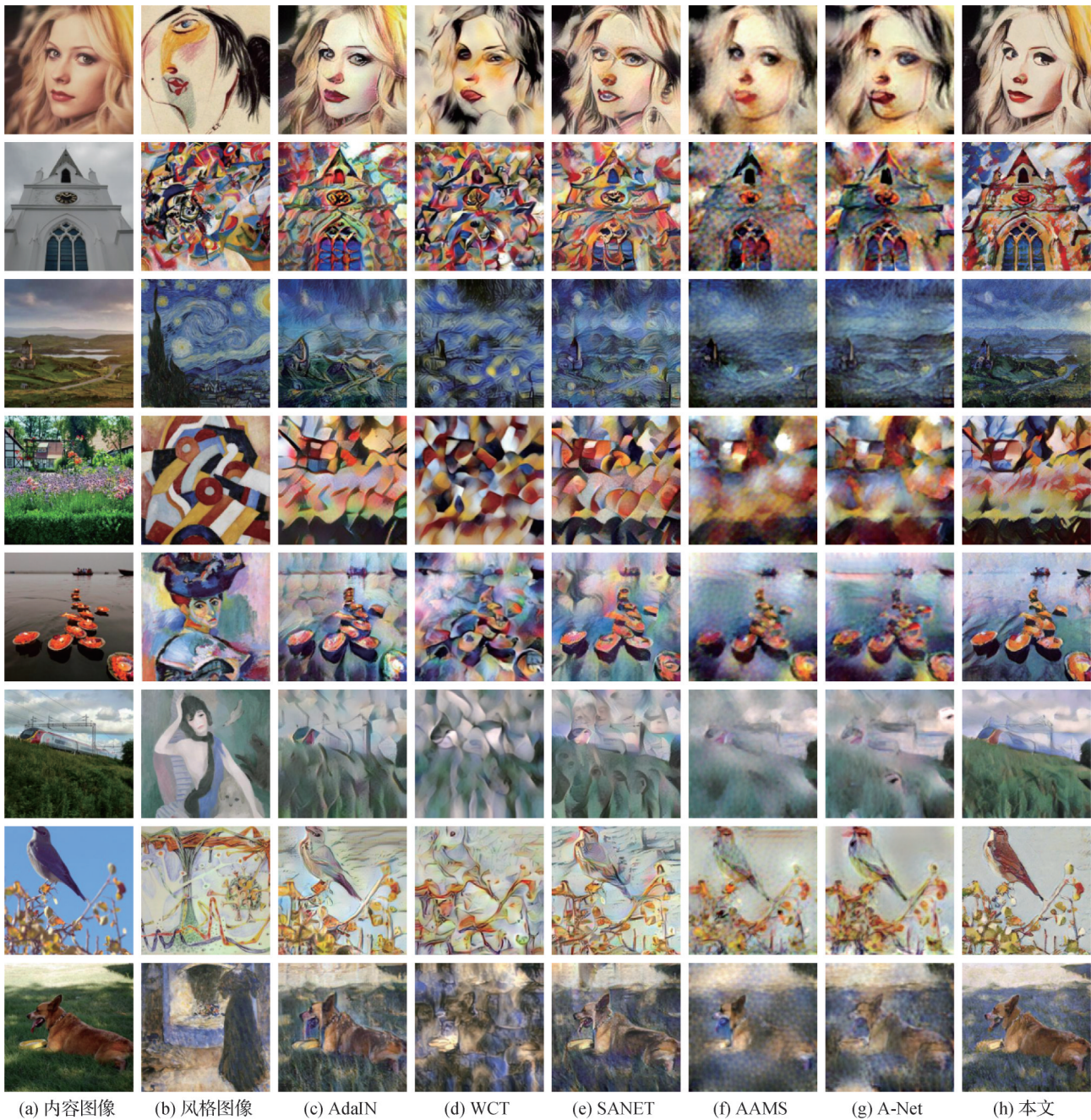


图5 与现有任意风格迁移方法生成效果进行比较

Fig. 5 Comparison the generated results with existing arbitrary style transfer methods
((a) content images;(b) style images;(c) AdaIN;(d) WCT;(e) SANET;(f) AAMS;(g) A-Net;(h) ours)

示,图像的结构发生了形变。直接使用风格图像块替换内容图像块能更好地引入风格笔触纹理, SANET使用注意力将内容特征块替换为风格特征块的加权和,生成结果较好地保留了结构和样式信息,但同时引入了不合理的风格内容结构,如图5(e)第1行图像中嘴巴部分。AAMS绘制多笔触风格化图像,画面整体较为平滑,但是受到区域聚类操作效果的影响,生成图像轮廓不够清晰,同时图像中主

体细粒度笔触在背景粗粒度笔触的映衬下略显突兀,如图5(f)所示。A-Net生成的图像较好地保留了风格特点,但是其结果依赖于纹理块划分的大小,当输入图像尺寸较小时,生成结果中内容轮廓有些难以分辨。与AdaIN和WCT相比,本文网络基于块替换方法,能更好地融合风格图像的样式特征,如图5(h)所示,生成的风格化图像与风格图像拥有相似的色彩表现。加之本文使用判别网络分辨风格图像的

真实性,拉近了生成结果与真实风格图像间的距离,提高了生成图像的光滑和细腻程度,如图 5(h) 第 1 行人像所示,线条流畅,脸部区域干净,无其他多余纹理,还有第 4、5 行图像的背景部分,都绘制得更加清爽,更好地突出了画面中的主体。

3.2.2 用户调查

通过用户调查,观察用户更喜欢哪一种任意风格迁移方法生成的风格化图像。随机选取网络生成的 15 幅风格化图像构建调查问卷,问卷中使用的图像为 256×256 像素。要求参与调查的用户综合考虑内容和风格因素,在每题中选择他们最喜欢的风格化图像,调查结果如图 6 所示。

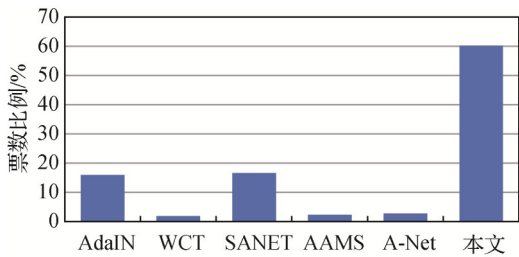


图 6 用户调查结果

Fig. 6 Survey results of users

3.2.3 速度比较

表 1 展示了各个任意风格迁移方法在 256×256 像素和 512×512 像素图像上进行风格化处理的平均时间。

所有方法模型均在相同实验环境下运行,表 1 中结果为使用单张 NVIDIA RTX A6000 GPU 运行 400 次风格迁移过程求取的平均风格化速率。本文方法相较 AdaIN 和 SANET 速度慢了 5~8 倍,但是相较于 WCT、AAMS 和 A-Net 方法风格化速度更快,因此,本文网络的风格化速率整体属于可接受范围内。当图像尺寸增至 512×512 像素时,风格化速度变化不大,表明了本文网络的稳定性。

3.2.4 与基于 Transformer 的风格迁移方法比较

Deng 等人(2022)使用 Transformer 网络捕获图像全局信息,解决内容信息泄露问题,同时提出更适用于图像风格迁移任务的内容感知位置编码(content-aware positional encoding, CAPE)方法,使得在图像尺寸变化时位置编码信息不变。本文方法与其不同之处在于:1)本文编、解码网络使用残差结构,引入图像低级纹理信息,增强图像特征的表

表 1 风格化速度比较

Table 1 Execution time comparison

方法	256 × 256 像素图像	512 × 512 像素图像
AdaIN	0.004	0.004
WCT	0.231	0.649
SANET	0.006	0.006
AAMS	0.573	1.371
A-Net	0.460	1.087
本文	0.031	0.033

注:加粗字体表示各列最优结果。

示能力;2)本文编、解码网络一同训练,使编码网络学习提取适合风格迁移的图像特征;3)本文使用判别网络度量生成图像,依此提高生成图像的真实性。

本文与 Deng 等人(2022)在 256 像素图像上的风格迁移效果如图 7 所示。可以看到, Deng 等人(2022)方法生成图像的纹理细节丰富、背景干净;本文方法生成图像具有风格图像的绘制笔触,如图 7(c)第 1 行的图像,具有油画质感,但是过多的笔触造成背景有些杂乱。



(a) 内容图像 (b) 风格图像 (c) 本文 (d) Deng 等人(2022)

图 7 与 Deng 等人(2022)方法进行比较

Fig. 7 Compare with Deng et al. (2022)((a) content images; (b) style images; (c) ours; (d) Deng et al. , (2022))

3.3 消融实验

本文设计消融实验以验证判别网络的有效性,实验结果如图 8 所示。

1)配置 1。去掉判别网络后,网络训练的基本配置包括内容损失函数和风格损失函数。从图 8(c)的风格化效果来看,生成图像保留了良好的内容结构,这得益于网络中残差结构引入了浅层表征信息。但是,图像上也出现了轻微的棋盘效应,画面中



图8 比较有无判别网络生成的风格化图像效果
Fig. 8 Compare the performance of stylized images generated with and without discriminative networks ((a) content images; (b) style images; (c) without discriminative networks; (d) with discriminative networks)

显示出规则性块状纹理;而从风格来看,尽管生成图像拥有目标风格分布,但是其风格笔触较为简单,例如当参考风格为油画类型时,生成图像过于光滑,没有油画特有的笔触信息,风格特点不够突出。

2)配置2。引入判别网络后,在配置1的基础上,增加对抗损失函数对网络训练过程进行约束,图8(d)展示了该配置下图像的风格化效果。可以看出,生成图像有了更明显的结构抽象效果,选择性地忽略了一些细节表现,降低了大面积区域多余纹理的出现,提升了画面渲染的简洁度和平滑度,更好地突出了画面主体部分。当参考图像为油画风格时,能够体现油画笔触特征,使其看起来更像画家创作的艺术作品。

3.4 其他应用方向

本文提出的任意风格迁移网络可以在不修改训练阶段网络结构的情况下完成其他的风格转移任务。

3.4.1 内容—风格权衡

除了在训练时调整内容损失函数和风格损失函数权重比控制风格化程度,训练好的网络在测试阶段也能够实现任意程度的风格化渲染效果。使用风格化权重 α 控制内容重建特征 f_{cc} 和风格化特征 f_{cs} 各自所占比例,实现风格化程度调整。具体为

$$Control(f_{cc}, f_{cs}, \alpha) = \alpha \times f_{cs} + (1 - \alpha) \times f_{cc} \quad (15)$$

式中,Control表示内容—风格的权衡操作, f_{cc} 是将两幅相同内容图像输入网络生成的内容重建特征, α 表示权衡比,取值范围为 $[0, 1]$ 。当 $\alpha = 0$ 时,网络生成原始内容图像 I_c ,当 $\alpha = 1$ 时,网络生成标准风格化图像 I_{cs} 。随着 α 值的增加,生成图像的风格特征越来越明显,与风格图像特征越来越相似,如图9所示。

3.4.2 风格插值

混合风格特征由不同的风格化特征经过加权和得到。设在 M 幅风格图像间进行风格插值,风格图像集合表示为 $\{I_s^m\}_{m=1}^M$,其中 I_s^m 表示集合中第 m 幅风格图像。风格图像对应的权重集合为 $\{w_m\}_{m=1}^M$,其中 w_m 表示第 m 幅风格图像所占权重,权重集合满足 $(\sum_{m=1}^M w_m) = 1$ 。依据风格集合可以获得风格化特征集合 $\{f_{cls_m}\}_{m=1}^M$,其中 f_{cls_m} 是由内容图像和第 m 幅风格图像生成的风格化特征,通过将风格化特征集合与权重集合对应元素相乘并求和,可以得到风格插值特

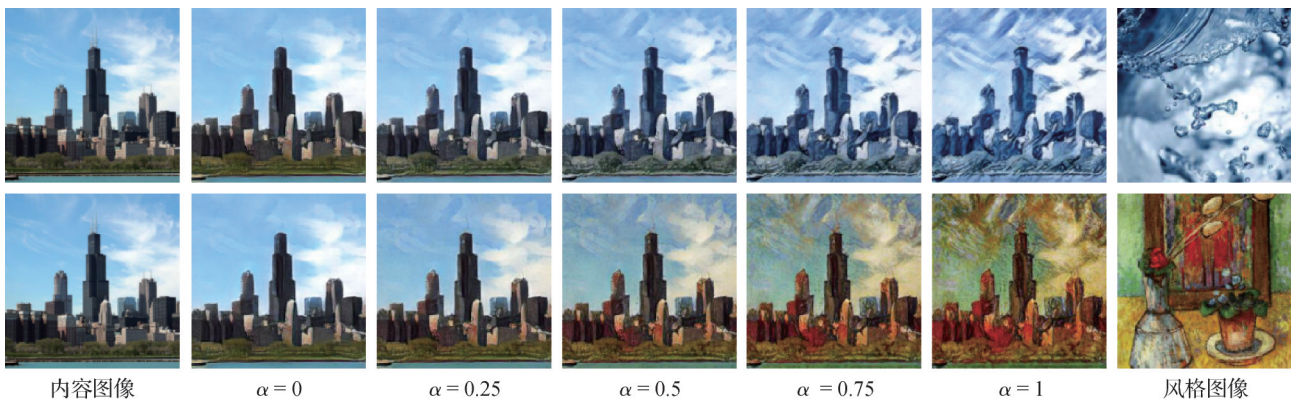


图9 控制权衡参数实现不同程度的风格化效果
Fig. 9 Control trade-off parameters to achieve different levels of stylized performance

征 f_{mix} 。具体为

$$f_{\text{mix}} = \sum_{m=1}^M f_{\text{cls}_m} w_m \quad (16)$$

以在两幅风格图像间进行风格插值为例,生成

的风格化图像如图10所示,图中 $s_1:s_2$ 形式的数据用来计算风格插值权重,计算过程为

$$w_1 = \frac{s_1}{s_1 + s_2}, w_2 = \frac{s_2}{s_1 + s_2} \quad (17)$$

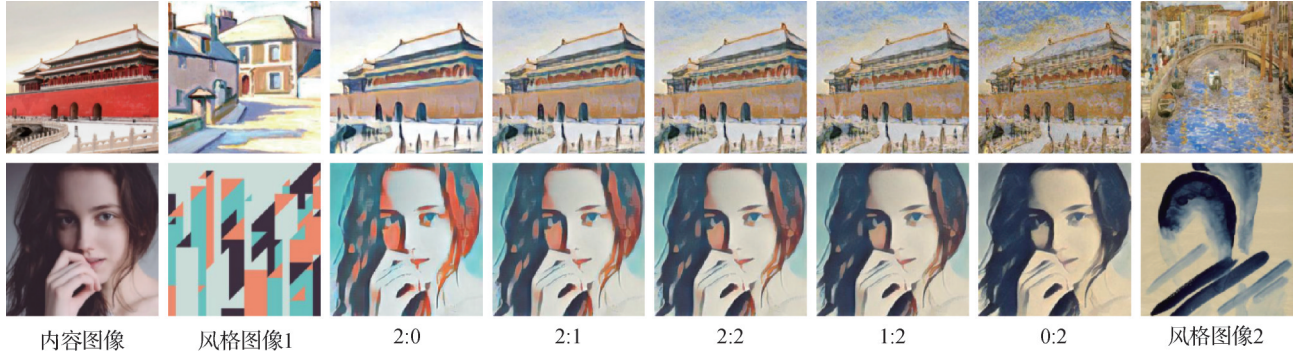


图10 使用不同插值权重生成的混合风格化效果

Fig. 10 Hybrid stylized results generated using different interpolation weights

3.4.3 区域绘制

本文模型不仅可以整体绘制风格化图像,还可以使用不同的风格特征绘制内容图像中的不同区域,即

$$f_{\text{area}} = \text{mask} \odot f_{\text{cls}_1} + (1 - \text{mask}) \odot f_{\text{cls}_2} \quad (18)$$

式中, f_{cls_1} 为使用内容图像 I_c 和风格图像 I_{s1} 生成的风格化特征, f_{cls_2} 表示使用 I_c 和风格图像 I_{s2} 生成的风格化特征, mask 为输入的二值图像,用以将内容图像划分为不同的区域, $\odot$ 表示元素间点乘操作。区域绘制效果如图11所示。可以看到,生成结果中人像主体与背景拥有不同的样式风格,但是二者在边界处融合自然,提高了图像整体的表现力。



图11 使用不同风格特征绘制内容图像不同区域的生成效果

Fig. 11 Results for drawing different areas of content images using different style features

4 结论

鉴于风格转换过程与句子翻译过程间的相似

性,本文提出一种结合卷积神经网络特征提取能力和Transformer网络全局依赖性捕获能力的混合网络模型,用于图像的任意风格迁移任务。首先,在卷积网络中引入残差连接,为提取的图像高级特征增添低级纹理信息,提高特征的表示能力和网络保留物体清晰结构的能力;其次,使用Transformer优秀的相关性建模能力,为特征增添全局信息,同时完成语义感知的特征块替换过程;针对风格信息难以用数值精确表示的问题,引入判别网络度量风格化效果。实验结果表明,本文网络相较其他方法能高效地生成高质量的风格化图像。

尽管本文网络能够实现良好的风格化效果,但仍存在一些问题需要完善。如,判别网络将风格判别过程视为一个二分类问题,即生成图像和真实风格图像,因此判别网络学习的是两个图像域之间的差异。如何调整判别网络使其能判别任意风格特征是后续研究工作的重要方向。

参考文献(References)

- Chen H B, Zhao L, Wang Z Z, Zhang H M, Zuo Z W, Li A L, Xing W and Lu D M. 2021a. Artistic style transfer with internal-external learning and contrastive learning//Proceedings of the 35th Conference on Neural Information Processing Systems. [s.l.]: [s.n.]
- Chen H T, Wang Y H, Guo T Y, Xu C, Deng Y P, Liu Z H, Ma S W, Xu C J, Xu C and Gao W. 2021b. Pre-trained image processing Transformer//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE;

- 12294-12305 [DOI: 10.1109/CVPR46437.2021.01212]
- Chen T Q and Schmidt M. 2016. Fast patch-based style transfer of arbitrary style [EB/OL]. [2021-12-26].
<https://arxiv.org/pdf/1612.04337v1.pdf>
- Chen Y, Lai Y K and Liu Y J. 2018. CartoonGAN: generative adversarial networks for photo cartoonization//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9465-9474 [DOI: 10.1109/CVPR.2018.00986]
- Deng Y Y, Tang F, Dong W M, Huang H B, Ma C Y and Xu C S. 2021. Arbitrary video style transfer via multi-channel correlation//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press: 1210-1217
- Deng Y Y, Tang F, Dong W M, Ma C Y, Pan X J, Wang L and Xu C S. 2022. StyTr²: image style transfer with Transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11316-11326 [DOI: 10.1109/CVPR52688.2022.01104]
- Deng Y Y, Tang F, Dong W M, Sun W, Huang F Y and Xu C S. 2020. Arbitrary style transfer via multi-adaptation network//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 2719-2727 [DOI: 10.1145/3394171.3414015]
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Housley N. 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. Virtual: ICLR
- Efros A A and Freeman W T. 2001. Image quilting for texture synthesis and transfer//Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles, USA: ACM: 341-346 [DOI: 10.1145/383259.383296]
- Esser P, Rombach R and Ommer B. 2021. Taming Transformers for high-resolution image synthesis//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12868-12878 [DOI: 10.1109/CVPR46437.2021.01268]
- Fu J, Liu J, Tian H J, Li Y, Bao Y J, Fang Z W and Lu H Q. 2019. Dual attention network for scene segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3141-3149 [DOI: 10.1109/CVPR.2019.00326]
- Gatys L A, Ecker A S and Bethge M. 2016. Image style transfer using convolutional neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2414-2423 [DOI: 10.1109/CVPR.2016.265]
- Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680
- Haeberli P. 1990. Paint by numbers: abstract image representations//Proceedings of the 17th Annual Conference on Computer Graphics and Interactive Techniques. Dallas, USA: ACM: 207-214 [DOI: 10.1145/97879.97902]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]
- Hertzmann A. 1998. Painterly rendering with curved brush strokes of multiple sizes//Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques. Orlando, USA: ACM: 453-460 [DOI: 10.1145/280814.280951]
- Hertzmann A, Jacobs C E, Oliver N, Curless B and Salesin D H. 2001. Image analogies//Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles, USA: ACM: 327-340 [DOI: 10.1145/383259.383295]
- Hicsonmez S, Samet N, Akbas E and Duygulu P. 2020. GANILLA: generative adversarial networks for image to illustration translation. Image and Vision Computing, 95: #103886 [DOI: 10.1016/j.imavis.2020.103886]
- Huang X and Belongie S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1510-1519 [DOI: 10.1109/ICCV.2017.167]
- Jiang Y F, Chang S Y and Wang Z Y. 2021. TransGAN: two Transformers can make one strong GAN [EB/OL]. [2021-12-26].
<https://arxiv.org/pdf/2102.07074v1.pdf>
- Johnson J, Alahi A and Li F F. 2016. Perceptual losses for real-time style transfer and super-resolution//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 694-711 [DOI: 10.1007/978-3-319-46475-6_43]
- Kotovenko D, Sanakoyeu A, Ma P C, Lang S and Ommer B. 2019. A content transformation block for image style transfer//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10024-10033 [DOI: 10.1109/CVPR.2019.01027]
- Lee H J, Kim H E and Nam H. 2019. SRM: a style-based recalibration module for convolutional neural networks//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1854-1862 [DOI: 10.1109/ICCV.2019.00194]
- Li Y J, Fang C, Yang J M, Wang Z W, Lu X and Yang M H. 2017. Universal style transfer via feature transforms//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 385-395
- Liu H C, Ren W Q, Wang R and Cao X C. 2022. A super-resolution Transformer fusion network for single blurred image. Journal of

- Image and Graphics, 27(5): 1616-1631 (刘花成, 任文琦, 王蕊, 操晓春. 2022. 用于单幅模糊图像超分辨率的Transformer融合网络. 中国图象图形学报, 27(5): 1616-1631) [DOI: 10.11834/jig.210847]
- Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin Transformer: hierarchical vision Transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 10012-10022
- Luan F J, Paris S, Shechtman E and Bala K. 2017. Deep photo style transfer//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6997-7005 [DOI: 10.1109/CVPR.2017.740]
- Luan F J, Paris S, Shechtman E and Bala K. 2018. Deep painterly harmonization. Computer Graphics Forum, 37(4): 95-106 [DOI: 10.1111/cgf.13478]
- Mirza M and Osindero S. 2014. Conditional generative adversarial nets [EB/OL]. [2021-12-26]. <https://arxiv.org/pdf/1411.1784v1.pdf>
- Park D Y and Lee K H. 2019. Arbitrary style transfer with style-attentional networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5873-5881 [DOI: 10.1109/CVPR.2019.00603]
- Parmar N, Vaswani A, Uszkoreit J, Kaiser L, Shazeer N, Ku A and Tran D. 2018. Image Transformer [EB/OL]. [2021-12-26]. <https://arxiv.org/pdf/1802.05751v3.pdf>
- Sheng L, Lin Z Y, Shao J and Wang X G. 2018. Avatar-Net: multi-scale zero-shot style transfer by feature decoration//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8242-8250 [DOI: 10.1109/CVPR.2018.00860]
- Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR: 1-12
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010
- Wang Z Z, Zhao L, Chen H B, Qiu L H, Mo Q H, Lin S H, Xing W and Lu D M. 2020. Diversified arbitrary style transfer via deep feature perturbation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7786-7795 [DOI: 10.1109/CVPR42600.2020.00781]
- Winnemöller H, Olsen S C and Gooch B. 2006. Real-time video abstraction//Proceedings of ACM SIGGRAPH 2006. Boston, USA: ACM: 1221-1226 [DOI: 10.1145/1179352.1142018]
- Xie B, Wang N and Fan Y W. 2020. Correlation alignment total variation model and algorithm for style transfer. Journal of Image and Graphics, 25(2): 241-254 (谢斌, 汪宁, 范有伟. 2020. 相关对齐的总变分风格迁移新模型. 中国图象图形学报, 25(2): 241-254) [DOI: 10.11834/jig.190199]
- Yanai K. 2017. Unseen style transfer based on a conditional fast style transfer network//Workshop of 5th International Conference on Learning Representations. Toulon, France: penReview.net
- Yao Y, Ren J Q, Xie X S, Liu W D, Liu Y J and Wang J. 2019. Attention-aware multi-stroke style transfer//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1467-1475 [DOI: 10.1109/CVPR.2019.00156]
- Zhang H, Goodfellow I, Metaxas D and Odena A. 2019. Self-attention generative adversarial networks//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 7354-7363.
- Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]

作者简介

孙梅婷, 女, 硕士研究生, 主要研究方向为图像风格迁移。

E-mail: sunmeiting@njust.edu.cn

代龙泉, 通信作者, 男, 副教授, 主要研究方向为图像处理和计算机视觉。E-mail: dailongquan@njust.edu.cn

唐金辉, 男, 教授, 主要研究方向为多媒体分析与检索、社交媒体分析、计算机视觉和人工智能。

E-mail: jinhuitang@njust.edu.cn