

基于风格特征分布匹配的图像风格迁移

米志鹏

安徽工业大学 计算机科学与技术学院, 安徽 马鞍山 243000

摘要:本研究针对现有方法在实现风格迁移任务时只能提取图像特征的低阶统计量这一问题,考虑将风格迁移过程建模为一个特征分布匹配过程,提出了一个基于 Wasserstein 距离的判别器网络并以此定义了一个风格损失函数,Wasserstein 判别器能够更好地拟合特征分布之间的 Wasserstein 距离,定义的风格损失也能够更好地区分图像特征的高阶统计信息之间的差异。同时,为了达到实时生成的效果,引入一个基于编码器-解码器结构和一个基于注意力机制的风格迁移转换模块作为生成网络,该生成网络能够有效融合原始图像特征并生成。具体而言,通过在计算损失模块的卷积层(CNN)后面添加 Wasserstein 判别器来计算风格损失,然后将风格损失与传统方法中计算为均方误差的内容损失一起监督生成网络的训练,在网络训练结束后,可以输入任意图像进行风格迁移测试。最后,在基准 MSCOCO 和 WikiArt 数据集上训练网络并测试结果,定性实验和定量实验结果表明,与现有方法相比,所提出的方法可以实现实时风格迁移,并且生成高质量风格化效果。

关键词:注意力机制; Wasserstein 距离; 特征分布; 风格迁移

中图分类号:TP183 **文献标识码:**A **doi:**10.16055/j.issn.1672-058X.2023.0002.008

Image Style Transfer Based on the Distribution Matching of the Style Features

MI Zhipeng

School of Computer Science and Technology, Anhui University of Technology, Anhui Maanshan 243000, China

Abstract: Aiming at the problem that the existing methods can only extract the low-order statistics of image features when implementing the task of style transfer, this study considered modeling the style transfer process as a feature distribution matching process, proposed a discriminator network based on Wasserstein distance, and defined a style loss function. Wasserstein discriminator can better fit the Wasserstein distance between feature distributions, and the defined style loss can better distinguish the difference between higher-order statistical information of image features. At the same time, in order to achieve the effect of real-time generation, a style transfer conversion module based on an encoder-decoder structure and an attention mechanism was introduced as a generation network. This generation network can effectively integrate the original image features and generate them. Specifically, the style loss was computed by adding a Wasserstein discriminator after the convolutional layer of the computational loss module. The training of the generative network was then supervised by the style loss together with the content loss calculated as the mean square error in the traditional methods, and after the network training, images can be input for the style transfer test. Finally, the network was trained on the benchmark MSCOCO and WikiArt datasets and the results were tested. The qualitative and quantitative experimental results showed that the proposed method can achieve real-time style transfer and generate high-quality stylization results compared with existing methods.

Keywords: attention mechanism; Wasserstein distance; feature distributions; style transfer

收稿日期:2021-11-24 **修回日期:**2021-12-26 **文章编号:**1672-058X(2023)02-0051-06

基金项目:安徽省自然科学基金资助(2008085MF190)。

作者简介:米志鹏(1997—),男,河北张家口人,硕士研究生,从事计算机视觉研究。

引用格式:米志鹏. 基于风格特征分布匹配的图像风格迁移[J]. 重庆工商大学学报(自然科学版), 2023, 40(2): 51—56.

MI Zhipeng. Image style transfer based on the distribution matching of the style features[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(2): 51—56.

1 引言

随着深度学习技术的快速发展以及各种深度模型在计算机视觉领域的广泛应用,深度学习图像生成的工作也逐渐涌现出来。图像风格迁移是当前数字艺术产业中一种有效的艺术图像生成技术,可广泛应用于艺术创作、数字媒体和移动应用。风格迁移的实现一般需要先获取艺术风格图像的纹理、颜色、形状等风格图案,然后将其传递到原始内容图像中,生成新的风格化图像。早期的风格迁移作品主要通过构建数据驱动的光照模型或合成纹理,使用不同的明暗处理或绘画风格来渲染摄影图像。近年来,由于在自动特征表示和提取方面取得的巨大成功,许多研究人员利用神经网络(DNN)^[1]来实现图像的转换和风格化。

当前,基于深度学习的图像风格迁移算法主要分为三大类:(1)基于迭代优化的方法^[2-3]:这些方法通常是给定一个原始噪声图像,通过定义的损失函数来反向优化噪声图像,最终生成图像达到一个平衡状态,既能具有原始内容图像的结构,同时具有风格图像的纹理、颜色等信息。然而,这些方法每次生成都需要一次较长的迭代过程,无法实现任意实时生成的效果,难以应用到实际场景当中。(2)又有基于前向生成网络的方法^[4]:该方法通过一个前向生成网络来实现风格迁移,能够产生实时的效果,然而,由于模型训练仅仅针对单一的风格,所以一旦网络训练完成,就可以使用任意的内容图像产生对应风格风格化图像。然而,由于网络模型的限制,当需要生成不同风格时,需要根据对应的风格图像重新训练网络,所以不具备任意风格生成的条件。(3)任意的风格迁移方法^[5-9]:这些方法能够实现任意实时的生成,通常要求一个生成网络来进行前向生成,在大规模的内容数据集和风格数据集上进行训练。一旦训练完成,输入任意的一对内容和风格图像,即可实现任意风格迁移,同时生成质量也非常好。然而,在这些方法中,通常把内容定义为图像的语义信息,而风格定义为图像的纹理信息(粗糙,平滑,颜色等)^[2]。通过定义内容损失和风格损失来驱动网络模型的学习,内容通常计算为特征图之间的均方误差(Mean Squared Error, MSE),而风格通常采用特征的偏心协方差矩阵(Gram Matrix)来表示。然而,这种计算依赖CNN提取风格特征的一阶或者二阶统计量,难以捕获更高级别的纹理,所以并不是最优的。

风格迁移中“何为风格”,一直没有明确的定义,Li等^[10]通过理论以及实验论证了风格可以表示为特征分布,并且风格损失可以定义为风格图像与生成的风格化图像之间的特征分布距离,但他们仍然基于一阶或者二阶统计来进行分布匹配,所以难以完全区分两个

概率分布。在最优传输理论中,根据深度学习的流形分布定律:一类自然数据可以被视作高维背景空间中的低维数据流形上的一个概率分布。深度学习的主要任务包括两个:学习流形结构和表示概率分布。在风格迁移中,本质上是学习原始风格和生成风格之间的最优传输映射关系^[11]。因此,对于原始特征分布和目标特征分布,Wasserstein距离成为更具表现力的度量,当然,此处存在的一个不足就是计算负担会有相应的增加。

一种有效的风格迁移方法通常需要生成的风格化图像既能保持原有内容图像的结构信息,同时把风格图像的风格图案迁移上去。文献[3]提出了基于松弛最优传输和自相似性(STROTSS)的风格迁移方法,这是一种新的基于优化的风格迁移算法,但是该方法不能实现实时图像生成。在最近的很多研究中,编码器-解码器的深度模型通常被用作风格迁移的主干网络。例如:文献[5]提出了一个新的自适应实例归一化(AdaIN)层,它将内容特征的均值和方差与风格特征的均值和方差对齐。文献[6]通过执行一对特征变换(白化和着色)来调整内容图像。但是,它可能导致样式特征扭曲,并可能在生成的结果中生成不需要的图案。文献[7]使用基于块的风格装饰器将内容特征映射到对应风格图案的特征,同时保持内容结构。然而,由于风格图像的风格图案的规模取决于块大小,因此很难平衡局部和全局风格模式。文献[8](SANet)采用自注意力结构学习从风格特征到内容特征的相关映射。文献[9]将内容相关注意力结构和风格相关注意力结构相结合,以分离内容和风格表示,并根据内容表示分布重新排列风格表示的分布。然而,以上方法的损失函数捕获的都是原始特征的一阶或者二阶统计量,并没有考虑到风格特征的分布度量,所以不能完全地提取风格信息。

基于上述研究,本文探索了使用编码器-解码器以及基于自注意力机制的转换模块作为主干生成网络来完成风格迁移的方法。该生成网络通过内容图像与风格图像的相关注意力来控制风格生成从而合成高质量的风格化图像,网络一旦训练完成,在测试阶段,输入一对任意的内容-风格图像,即可实现任意实时的风格迁移。同时,基于最近W-GAN在图像生成任务当中体现出的先进性,Wasserstein距离在计算相似性时相较于欧式距离等具有先进性。在Arjovsky^[12]的论证中,分析并证明了它相比JS散度(Jensen-Shannon Divergence),KL散度(Kullback-Leibler Divergence)和总变分(Total Variation)更具有优越性。基于此,引入了风格判别器网络,该判别器网络近似原始特征分布和生成特征分布之间的Wasserstein距离,并以此定义了

风格损失函数,该风格损失函数能够针对相应的特征分布计算更高阶的特征统计信息,相较于传统方法中定义的风格损失函数具备更好的可靠性。最后,将定义的 Wasserstein 风格损失与传统的内容损失一起作为整个模型的损失函数来监督训练。

综上,本方法在构建一个高效的风格迁移主干网络的基础上,通过引入一个基于最优传输理论的 Wasserstein 判别器,并基于 Wasserstein 距离定义近似原始风格图像特征分布与生成图像特征分布之间的风格损失函数。当网络模型训练结束后,在测试阶段,选取任意的一对内容-风格图像作为模型输入,能够实时生成高质量的风格化图像,生成图像在不损坏原始内容图像结构信息的同时能够有效地迁移风格图案,实验结果证明了本文提出方法的有效性。

2 网络模型架构

2.1 网络模型总体结构

图 1 展示了深度网络的总体结构图,输入为内容图像 I_c 与风格图像 I_s ,首先经过编码器网络获取图像特征图 f_c 和 f_s ,然后将特征图输入到基于自注意力的风格转换块中完成特征融合得到 f_{cs} ,最后将融合特征图 f_{cs} 输入到解码器中得到输出图像 I_{cs} 。训练期间,内容损失计算为原始内容图像与生成图像特征图之间的 MSE 损失;风格损失通过引入的 Wasserstein 判别器进行计算,内容损失和风格损失共同监督网络的训练。

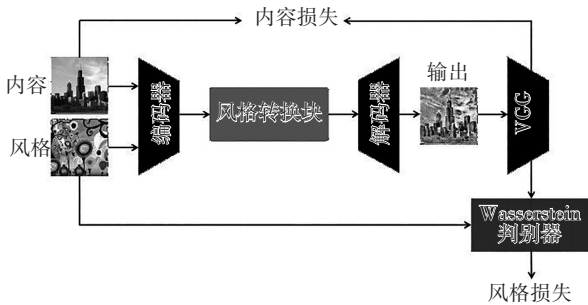


图 1 网络结构图

Fig. 1 Structure diagram of the proposed network

2.2 主干网络

主干网络的整体框架是编码器-解码器结构,其中编码器采用预训练好的 VGG-19^[1] 网络。风格转换块的核心是自注意力机制^[14],它关注所有位置并在嵌入空间中取其加权平均值来计算序列或图像中某个位置的响应。所采用的风格转换块通过稍微修改自我注意力机制来学习内容特征和风格特征之间的映射。

首先使用编码器来提取原始图像特征,如式(1),式(2)所示:

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

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

其中, Enc 表示 VGG 编码器, I_c 和 I_s 代表输入的内容图像和风格图像, f_c 和 f_s 代表编码后的内容特征图和风格图。

然后,将获得的特征图输入到风格转换块来完成特征融合,如式(3)所示:

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

其中, f_{cs} 代表融合后的风格化特征, SA 代表稍加修改的自注意力机制,具体如下:

$$Q_c = \text{Conv}(\tilde{f}_c)$$

$$K_s = \text{Conv}(\tilde{f}_s)$$

$$V_s = \text{Conv}(f_s)$$

$$\text{Att}_c = \text{Softmax}(Q_c \otimes K_s)$$

$$f_{cs} = \text{Att}_c \otimes V_s + f_c$$

其中, Q_c 代表对内容特征求 Query, K_s 代表对风格特征求 Key, V_s 代表对风格特征求 Value,它们的实现是通过

1×1 大小的卷积 (Conv)。 $\tilde{f}_c$ 代表沿着通道归一化后的内容特征图, $\tilde{f}_s$ 代表沿着通道归一化后的风格特征图。 Att_c 代表根据计算得到的注意力分数。 $\otimes$ 表示矩阵乘法。

最后,如式(4)所示,将得到的融合特征图输入解码器网络 (Dec) 即可得到最终的风格化图像 I_{cs} :

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

3 损失函数

本节主要介绍网络训练的损失函数,其中包括传统的内容损失 L_c (式(5)) 以及提出的基于 Wasserstein 判别器计算出的风格损失 L_s (式(6))。

3.1 内容损失

内容损失定义如式(5),采用的是常用的感知损失:使用预训练的 VGG-19 (Enc) 提取图像特征图,然后计算原始内容图像与生成图像特征图之间的 MSE 损失。

$$L_c = \| \text{Enc}(f_{cs}) - \text{Enc}(f_c) \|_2 \quad (5)$$

3.2 Wasserstein 距离

直观上来看, Wasserstein 距离可以解释为将一个质量分布移动到目标位置所需的最小工作量^[16]。这就是为什么它有时被称为“推土机距离”(Earth-Mover distance)。假设 p, q 是两个概率分布, Wasserstein 距离定义如下:

$$W(p, q) = \inf_{\gamma \in \Gamma(p, q)} E_{x, y \sim \gamma} [d(x, y)]$$

$\gamma \in \Gamma(p, q)$ 是 p, q 的所有联合概率分布, d 是某种距离函数(例如欧几里得距离)。Wasserstein 可以区分两个概率分布:对任意的 $\gamma \in \Gamma(p, q)$, 存在 x, y 使得 $x \neq y$ 且 $\gamma(x, y) > 0$ 。因为 $x \neq y, d(x, y) > 0$, 所以 $W(p, q) > 0$ 。

Arjovsky 等^[12]将 Wasserstein 距离用作生成性对抗

网络(W-GAN)中的生成器损失。在W-GAN中,判别器是一个经过训练的神经网络,用于近似模拟来自生成器的假数据分布和真实数据分布之间的 Wasserstein 距离,Arjovsky 等人通过理论和实验论证了 Wasserstein 距离的优越性。

Gulrajani 等^[13]后来改进了W-GAN的近似方法,改进的方法称为 Wasserstein 梯度惩罚(W-GAN-GP)。本文中定义的基于判别器的风格损失就是基于该方法来近似原始风格图像特征分布和生成风格化图像特征分布之间的 Wasserstein 距离。

3.3 Wasserstein 判别器

首先,引入的判别器作用是区分两个概率分布,这里判别器有3个隐藏层,尺寸为256,以及一个ReLU激活层。判别器的目的是拟合两个分布之间的 Wasserstein 距离,在本文实验中源分布和目标分布就是原始风格特征与生成的风格特征,示意图如图2所示。

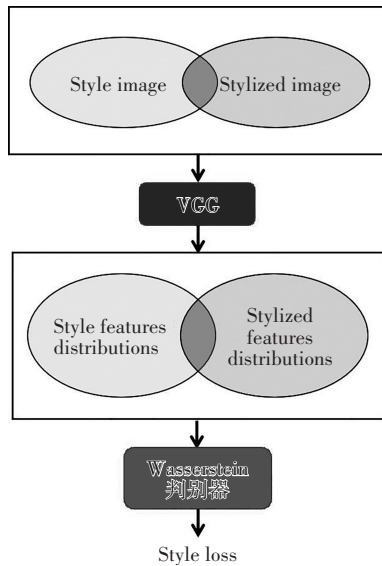


图2 基于 Wasserstein 判别计算风格损失

Fig. 2 The style loss based on Wasserstein discrimination

3.4 风格损失

为了获取需要计算的特征分布,首先需要得到VGG-19中间层输出特征图,此处取:conv1_1,conv2_1,conv3_1,conv4_1,conv5_1。为了使用 Wasserstein 距离计算风格损失,要在相应的CNN层上附加上一个鉴别器网络。这些鉴别器近似原始风格特征和W-GAN-GP下生成的特征之间的 Wasserstein 距离。风格损失 L_s 定义过程如下:

$$L_s^l = Dis[f_{cs}^l, f_s^l]$$

$$L_s = \sum_l w_l L_s^l \quad (6)$$

其中, Dis 表示 Wasserstein 判别器, l 表示CNN的中间层, f_{cs}^l, f_s^l 分别表示第 l 层得到的生成风格化特征与第

l 层得到的原始风格特征。最后,将每一层得到的结果累加得到风格损失和, w_l 表示不同层对应的权重参数,最后得到 L_s 即为风格损失函数,如式(6)。

3.5 总损失

总损失表示为式(7): L_t 内容损失与风格损失的加权和,实验中, α 设置为1, β 设置为10;在训练期间,两个损失共同监督主干网络的训练,一旦训练结束,直接将输入图像输入到保存的模型中即可实现实时风格迁移。

$$L_t = \alpha L_c + \beta L_s \quad (7)$$

4 实验结果与分析

在本节中展示了定性以及定量实验结果的比较与分析。其中实验配置如下:训练所需的内容图像数据和风格图像数据集分别为 MSCOCO^[1]和 Wikiart^[2];实验的图形显卡平台为 Nvidia 2080TI GPU;训练期间的优化器为 Adam,初始学习率设置为 $10e-4$,学习率衰减策略为每个迭代之后降低 $10e-5$;批大小设置为2;输入训练图片首先重新调整大小为 $512 * 512$,然后随机裁剪为 $256 * 256$ 。

4.1 定性结果及分析

图3中展示了与SANet的定性比较结果:对于左侧的输入内容图像和风格图像,第二列为原始SANet生成的结果;第三列为引入的风格损失以及SANet主干网络生成的结果。实验结果显示,引入的 Wasserstein 判别器能够有效地监督主干网络的训练,训练完成后,测试结果与基准方法相比较,视觉上具有优越性;在第二行展示了细节的放大比较,可以发现,在相同条件下,提出的体系结构产生的结果不仅有效迁移了风格,同时减少了原始内容图像的扭曲以及伪影现象,建筑物的窗口依旧清晰可见,由此看出提出的方法能够比较好地提升图像表现力。



图3 与 SANet 的局部对比

Fig. 3 The local comparisons with SANet

然后进一步地与 SANet 进行了比较。由于原始 SANet 存在的一个鲜明的问题是生成图像容易具有伪影,一般认为是自注意力机制产生的影响,但是通过实验发现,引入的风格损失同样较好地减轻了这一效应。进一步的实验结果如图 4 所示,第三列为 SANet 产生的结果,第四列为提出的方法产生的结果。结果显示:提出的方法产生的结果更加平滑,颜色分布统一,纹理也较为丰富,并且相较于 SANet,较大程度减少了伪影,图像扭曲现象减弱,较好地保持了原有内容图像的结构信息。

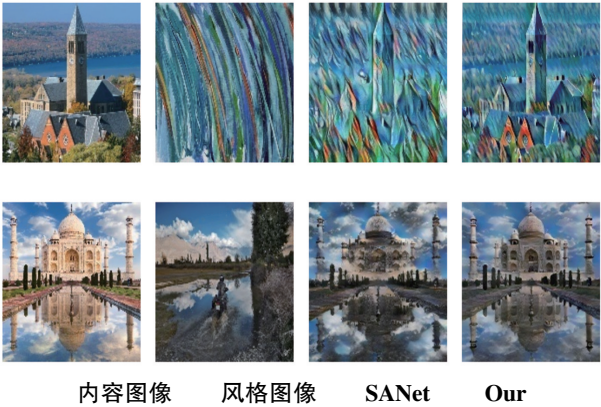


图 4 对比 SANet

如图 5 所示,对所提出的方法产生的结果与 AdaIN 产生的结果进行定性比较,AdaIN 自适应地将内容特征的均值和方差与风格特征的均值与方差进行对齐,速度快,无需训练相应的转换块,但是容易丢失原有图像细节,生成图像容易模糊。图中,第三列为 AdaIN 产生的结果,第四列为提出的方法产生的结果。结果显示:提出的方法生成的结果具有优越性,风格表现力较好,减少了失真。更多的实验结果比较请见图 6。



图 5 对比 AdaIN

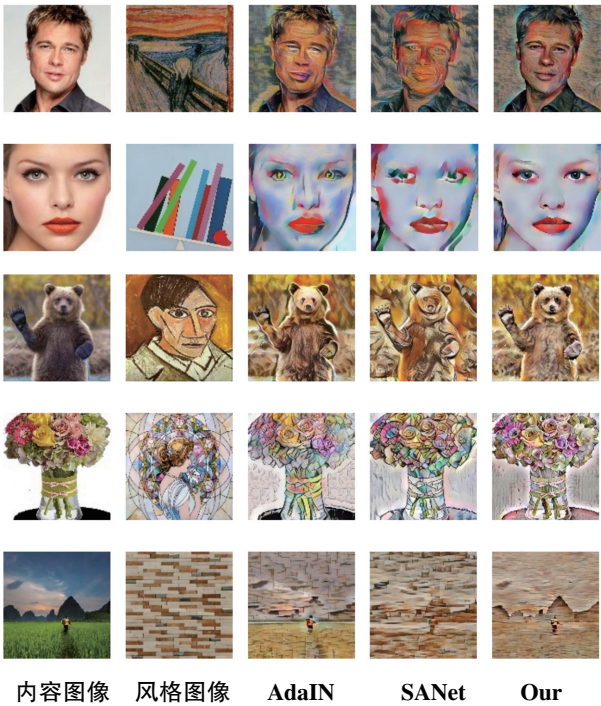


图 6 更多实验结果

4.2 定量结果及分析

为了更好地进行比较,这里引入了两个定量评估因子:User study(分值越高越好)与 Runtime(越低表示越快)。其中,User style 是风格迁移领域常用的一个主观用户评价指标,具有较好的说服力;Runtime 是输入图像到模型中产生结果所需时间。

对于 User study,采用的评估策略如下:首先随机选取 10 张内容图像和 10 张风格图像作为测试样本,将他们分别输入到要对比的方法模型中生成结果,每个方法产生 100 张对应的风格化图像。然后将相同输入使用不同方法生成的结果提供给 30 名用户进行视觉评分,评估指标为:对原始内容图像的保持程度;风格迁移的有效程度。用户根据这两个指标进行主观评分,分值在 0~1 之间。最后对所有的评分结果求平均值,结果如表 1 所示,提出的方法得到最高的用户分数。

表 1 定量分析
Table 1 Quantitative analysis

度量因子	方 法		
	AdaIN	SANet	Our
User study ↑	0.5	0.67	0.79
Runtime (s) ↓	0.038	0.078	0.039

对于 Runtime,同样随机选取 10 张内容图像和 10 张风格图像逐一输入到待测模型中,计算每个方法生成图像所需时间,最后总结归纳求出平均值。AdaIN 生成速度略快于本文提出的模型,原因是它仅仅采用特

征均值方差进行对齐,并不需要额外的模型参数,相对计算量较低,但是本文的方法能够在保持生成图像质量的情况下,生成速度也能够与其达到相当的程度。

5 结 论

本文研究了基于特征分布匹配的风格迁移算法,考虑了高阶统计特征分布匹配过程以及图像实时生成要素。提出的模型结构主要包括一个生成网络和一个基于 Wasserstein 判别器定义的风格损失函数,该损失函数与传统的内容损失函数共同监督整个生成网络的训练。实验中将本文方法生成的结果与现有的基准方法生成的结果进行了比较,实验结果显示了所提出的方法具有优越性,提出的方法生成的图像不仅能够保持原有内容图像的结构细节信息,同时能够有效地将风格图案迁移到内容图像。并且由于训练完成后,输入任意的内容风格图像到模型中,可以完成实时的迁移。因此,在实际应用中,有利于部署,满足现实场景的需求。

参考文献(References):

- [1] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014(14): 1—14.
- [2] GATYS L A, ECKER A S, BETHGE M. Image style transfer using convolutional neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2414—2423.
- [3] KOLKIN N, SALAVON J, SHAKHNAROVICH G. Style transfer by relaxed optimal transport and self-similarity [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nevada: IEEE, 2019, 10051—10060.
- [4] JOHNSON J, ALAHI A, FEI-FEI L. Perceptual losses for real-time style transfer and super-resolution [C]//European Conference on Computer Vision. Springer, Cham, 2016: 694—711.
- [5] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization [C]//Proceedings of the IEEE International Conference on Computer Vision. Hawaii: IEEE, 2017: 1501—1510.
- [6] LI Y, FANG C, YANG J, et al. Universal style transfer via feature transforms [J]. Advances in Neural Information Processing Systems, 2017: 386—396.
- [7] SHENG L, LIN Z, SHAO J, et al. Avatar-net: multi-scale zero-shot style transfer by feature decoration [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Utah: IEEE, 2018: 8242—8250.
- [8] 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. California: IEEE, 2019, 5880—5888.
- [9] DENG Y, TANG F, DONG W, et al. Arbitrary style transfer via multi-adaptation network [C]//Proceedings of the 28th ACM International Conference on Multimedia. Washington: ACM SIGMM, 2020, 2719—2727.
- [10] LI Y, WANG N, LIU J, et al. Demystifying neural style transfer [C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne: Morgan Kaufman, 2017, 2230—2236.
- [11] AN D, GUO Y, LEI N, et al. Ae-ot: a new generative model based on extended semi-discrete optimal transport [C]//International Conference on Learning Representations. Addis Ababa: ICLR, 2020, 19.
- [12] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks [C]//International Conference on Machine Learning. PMLR, 2017(70): 214—223.
- [13] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein gans [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 5769—5779.
- [14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. California: The MIT Press, 2017, 5998—6008.
- [15] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Utah: IEEE, 2018, 7794—7803.
- [16] HUANG E, GUPTA S. Style is a distribution of features [J]. arXiv preprint arXiv: 2007.13010, 2020.
- [17] MROUEH Y. Wasserstein style transfer [C]//International Conference on Artificial Intelligence and Statistics. PMLR, 2020: 842—852.
- [18] LIU H, GU X, SAMARAS D. Wasserstein gan with quadratic transport cost [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. California: IEEE, 2019, 4832—4841.
- [19] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context [C]//European Conference on Computer Vision. Springer, Cham, 2014: 740—755.
- [20] PHILLIPS F, MACKINTOSH B. WIKI ART GALLERY INC. A case for critical thinking [J]. Issues in Accounting Education, 2011, 26(3): 593—608.