基于字形约束和注意力的艺术字体风格迁移

吕文锐[1]，普园媛[1,2*]，赵征鹏[1]，张衡[1]，阳秋霞[1]

（1. 云南大学信息学院，云南 昆明 650504；2. 云南省高校物联网技术及应用重点实验室，云南 昆明 650504）

摘　要： 艺术字体的风格迁移是一项非常有趣但又十分具有挑战性的任务，具体来说就是将目标字体的艺术风格通过某种映射方式迁移到源字体上。现有方法在字形风格迁移方面存在鲁棒性有限的不足，且当 2 种不同风格的字形相差较大时不能很好地将风格内容迁移到目标字体上。针对以上问题，提出一种端到端的通用网络框架模型，并在模型中引入自注意力机制和自适应实例归一化，用于实现在给定的多个文本效果域之间进行任意字体的艺术风格迁移。该模型主要包括 1 个生成器和 2 个鉴别器，还有 1 个额外的风格编码器。为了更好地做到字形约束以及提升网络的性能，设计几种损失函数来优化生成对抗网络（GAN）的训练。为了验证该模型的有效性，采用了 FET-GAN 任务中公开的艺术字体数据集。实验对比了 6 种先进的方法，并从定量和定性 2 个方面进行了比较。实验结果表明，所提模型能够实现带有字体变换的字形图像风格迁移，迁移结果能够保持很好的字形结构，并且 FID 值为 72.355，低于对比实验中最好的结果 91.435。

关键词： 字体风格迁移；自注意力；自适应实例归一化；生成对抗网络；字形约束

源代码链接：https://pan.baidu.com/s/1UxrZc5AROMOSFtFPJDS9bw?pwd=v6uy

中图分类号：TP391　　**文献标志码：** A　　**DOI：** 10.19678/j.issn.1000-3428.0068380

Artistic Font Style Transfer Based on Glyph Constraints and Attention

LÜ Wenrui[1], PU Yuanyuan[1,2*], ZHAO Zhengpeng[1], ZHANG Heng[1], YANG Qiuxia[1]

(1. School of Information, Yunnan University, Kunming 650504, Yunnan, China;
2. Yunnan Province Universities Key Laboratory of Internet of Things Technology and Application, Kunming 650504, Yunnan, China)

【Abstract】 Art font style transfer is an intriguing yet challenging task that involves transferring the art style of a source font to a target font through mapping. This study aims to address the limitations of existing methods, particularly their limited robustness in font style migration and poor performance when there is a significant difference between the styles of the source and target fonts. To tackle these challenges, we proposed an end-to-end general network framework model incorporating a self-attention mechanism and adaptive instance normalization to realize artistic style transfer across multiple-text effect domains. The proposed model comprised a generator, two discriminators, and an additional style encoder. To better preserve the font structure and improve network performance, we designed several custom loss functions to optimize the training of a Generative Adversarial Network(GAN). The model was validated using a publicly available art font dataset for the FET-GAN task. In experiments comparing six state-of-the-art methods, the proposed method demonstrated superior performance both quantitatively and qualitatively. Extensive experimental results showed that the model effectively performs font image style migration while maintaining the glyph structure. The Fréchet Inception Distance(FID) of the proposed method is 72.355, which was notably lower than the best comparative result of 91.435, highlighting the method′s effectiveness.

【Key words】 font style transfer; self-attention; adaptive instance normalization; Generative Adversarial Network (GAN); the glyph constraints

0　引言

字体的艺术风格是文本的附加样式功能，比如颜色、字形轮廓、笔画、线条、发光和纹理等。以参考图像指定的风格来呈现文本内容的方式被称为风格迁移，而艺术字形图像风格迁移的目的就是以参考图像指定的艺术风格来呈现源字形图像的文本内容。到目前为止，已经有很多研究人员在艺术字体风格迁移领域开展了相关研究，其中开创性的神经网络风格迁移[1]的成功，掀起了基

收稿日期： 2023-09-12　　**修回日期：** 2024-01-03

基金项目： 国家自然科学基金（61271361，61761046，62162068，62362070）；云南省科技厅应用基础研究计划重点项目（202001BB050043）；云南省科技重大专项（202302AF080006）。

通信作者 E-mail： * yuanyuanpu@ynu.edu.cn

计算机工程 www.ecice06.com

于深度学习的图像风格化研究热潮，它的主要思想就是通过最小化 Gram 矩阵[2]之间的差异来匹配风格图像和生成的图像之间的全局特征分布[3]，但是这种用全局统计信息来表示通用样式的方法并不适用于文字风格迁移。因为字体的样式特征沿着字形高度结构化，所以不能简单地描述为均值、方差等其他统计量[4]。文献[5]的工作主要关注于字形几何轮廓的生成，其目的是找到字形轮廓之间精确的对应关系，但是当字形的数量和复杂度大大增加时，比如针对中文字体时，这种方法通常会失效。

最近，随着生成对抗网络(GAN)[6]的兴起，合成更逼真且更高质量的字形图像变得更加容易。但是，初期的 GAN 训练过程往往是不可控的，生成的图像经常会出现模糊或者伪影的现象。TET-GAN[7]将字体和视觉效果视为 2 个独立的属性，这种方法需要训练多个子网络以避免字体更换带来的干扰，而且不能进行字体的转换。

针对上述问题，本文提出了一种端到端的通用网络框架，用于实现带有字体变换的字形图像风格迁移。当给定一个属于同一风格域的参考图像时，该模型可以在保持全局特征不变的情况下将源图像的风格转换为参考图像的风格。本文为参考图像设计了一个风格编码器，用来提取参考图像的特征。此外，本文采用 GAN 的训练策略训练了一个生成器。它可以通过自适应实例归一化方法将源图像进行编码后在瓶颈层与编码器提取的风格特征进行融合，从而生成具有参考图像的风格，同时又能保持和源图像一致字形的图像。通过大量实验证明，本文的模型在字形图像的合成方面和最新的一些方法相比具有一定的优势。

总结来说，本文工作的主要贡献有以下 3 个方面：

1)本文在生成字体和目标字体之间采用了 HED(Holistically-nested Edge Detection)边缘检测，同时计算了相应的损失函数，使得网络在训练过程中能够更好地学习到目标字体的轮廓特征，对生成字体的字形进行约束，解决了以往方法生成字形的字体结构不完整的问题。

2)本文在生成器的上采样部分以及鉴别器的最后一个卷积层均采用了自注意力机制和光谱归一化，其目的是改善生成图像的纹理细节和增强网络训练的稳定性，一定程度上解决了生成图像质量差的问题。

3)本文采用了一个额外的局部鉴别器，并引入了一种计算高效的局部纹理细化损失函数，该损失函数有助于生成更好的局部纹理特征，有效解决了生成图像局部纹理细节丢失的问题。

1 相关工作

1.1 基于神经网络的风格迁移

图像风格迁移是将参考图像的风格样式迁移到内容图像的过程，这个过程与图像的纹理合成密切相关。GATYS 等[1]开创的卷积神经网络模型对图像的纹理合成就具有非常强大的表征能力。这种基于深度卷积神经网络的迁移算法打破了常规方法只能处理特定纹理风格的局面，并且其生成的结果也具有很强的语义对应性。但是这种方法严重依赖于优化过程，导致其训练过程缓慢。

后续很多工作在不同方面对其进行了改进。为了加快风格化训练速度，LI 等[8]试图训练只需通过一次前向传播就能实现风格化的前馈网络，但这种方法的局限在于每个网络仅适用于一种或几种指定的样式。为了兼顾灵活性与快速化，后续的研究进行了诸多尝试，比如文献[9]试图寻找一种通用的风格表示来替换 Gram 矩阵，其在迁移任务中使用了自适应实例归一化方法，通过该方法生成的图像具有更强的真实性和多样性，目前这一方法已被广泛应用到其他任务，比如无监督学习[10]和高质量的图像翻译[11]。在本文任务中也使用了类似的归一化方法，其目的在于将参考图像的风格编码映射为仿射参数，并将该仿射参数馈送到生成器中，和源图像的特征码进行融合，从而得到最终的生成图像。此外，WANG 等[12]深入探讨了基于补丁的样式转换的核心样式交换过程，并探索了使其多样化的方法。KALISCHEK 等[13]采用最近提出的用于域自适应的中心矩差异(CMD)的对偶形式，以实现目标样式和输出图像的特征分布之间差异的最小化。CANHAM 等[14]提出了一种基于视觉科学原理的图像和视频的照片级真实感风格迁移方法，他们提出了匹配功率谱来描述风格图片和目标图片之间的扩散效应，该扩散效应在以往都没有被考虑过。为了克服重复伪影的问题，HONG 等[15]专注于增强注意力机制并捕捉影响分割的主要因素，他们引入了样式可重复性，用来量化风格图像中样式的可重复性。文献[16]为了更好地保存和传递输入图像提取的特征，在生成器中加入了位置归一化和动态矩阵捷径模块，之后在重建图像中应用了多尺度结构相似性和 L_1 正则化来加强重建图像在对比度和

计算机工程 www.ecice06.com

结构方面的约束。文献[17]基于 CycleGAN 的基本框架进行优化改进，为了解决训练不稳定的问题，其将 CycleGAN 原判别器网络中批量归一化技术替换成梯度归一化技术，以此使得判别器在梯度空间更加平滑，同时还引入了新的 ResNeSt 残差网络，替代之前的 ResNet 残差网络。

这种新的方法可以更精确地匹配分布，从而更真实地再现所需的样式，同时仍然具有很高的计算效率。REN 等[18]处理了一项新任务，即文本引导图像操作，该方法实现了基于样式转换的图像操作，该方法不需要任何参考样式图像，直接实现将用户输入的句子生成样式图像。

1.2 文字风格迁移

文字是人们日常生活中最常见的视觉元素之一，早期有一些关于文字风格迁移方面的工作，比如文献[19]训练了卷积神经网络来学习笔画样式以进行字体迁移。但是，对带有艺术风格文字的研究并不多。在这一领域，文献[9]最早开始研究文字风格迁移问题，该文作者一开始根据字形笔画之间的标准化位置来匹配风格化的补丁，取得了令人满意的效果。但是由于需要通过大规模的检测来匹配最佳的补丁和距离估计，因此该方法需要耗费巨大的计算量。同时，字体的结构差异和样式特征对结果的影响也很大。文献[9]首次通过深度卷积神经网络来进行文字风格转换，利用 GAN，采用 3 个编码器和 2 个生成器组合成 3 对自动编码器架构，同时实现字形图像的样式分离和样式转换，但是不同的字形结构对结果的影响很大。文献[20]提出了一种基于层次结构的书法风格转换方法，该方法可以生成指定风格的书法风格。

总体而言，之前的大部分工作都只注重于视觉效果的传递而忽视了字体的变换。文献[21]提出了一个端到端的网络框架，并在字形转换任务中引入了自适应实例归一化和多类鉴别器，以实现在多个风格域之间带有字体变换的视觉效果传递。该文提出的模型在部分艺术字体之间的效果转换取得了令人满意的结果，但是泛化能力较差，对于纯文本字型和带有风格的艺术字体之间的转换效果并不理想。本文的模型采用端到端的网络框架来实现带有字体变化的文字风格迁移，并通过在目标图像和生成图像之间计算上下文损失来确保样式变换和字体迁移的准确性。

2 主要方法

2.1 模型网络框架

本文的目标是生成带有字体变换和风格转移的字形图像，因此设计了一个端到端的生成对抗网络。本文提出的算法框架如图 1 所示(彩色效果见《计算机工程》官网 HTML 版，下同)，该网络主要包含 1 个生成器和 2 个鉴别器，还有 1 个额外的风格编码器。为了更好地约束字形结构以及提升网络的性能，本文专门设计了几种损失函数来优化 GAN 的训练。

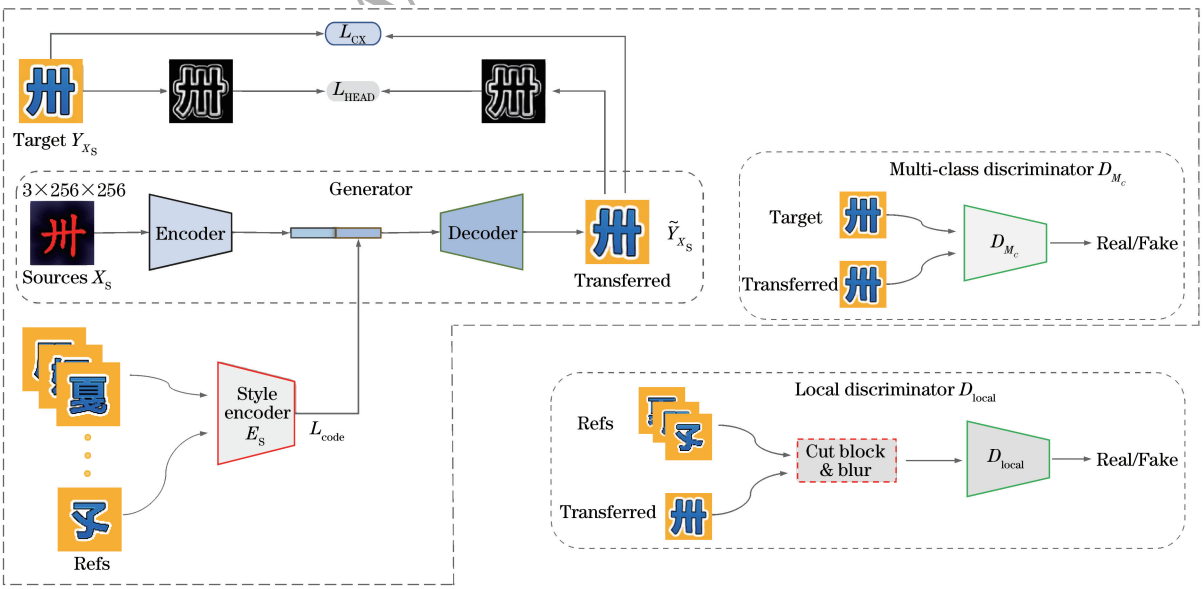


图 1 艺术字体风格迁移系统框架示意图

Fig. 1 Schematic diagram of the framework of artistic font style transfer system

本文模型能达到的风格迁移效果如图 2 所示。可以看出，给定 N 个参考图像，该模型可以生成和参考图像字体一致的字符图像，同时其字形和内容图像保持一致。

计算机工程 www.ecice06.com

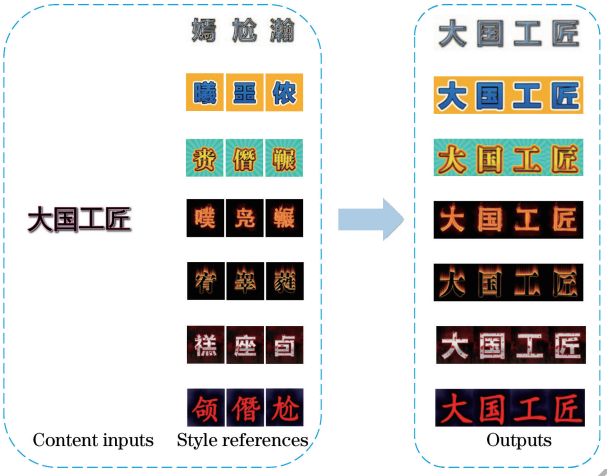


图 2 本文方法效果图

Fig. 2 Effect diagram of the proposed method

如图 2 所示，本文训练了一个 GAN 来执行艺术字风格迁移的任务，该模型包括的 2 个判别器，分别是多分类判别器 D_{M_C} 和局部判别器 D_{local}。多分类判别器的输出通道为训练集的类别数 C，对于多分类判别器，每次输入一类风格图片（共有 C 类）和一张与该风格对应迁移后得到的图片来判别真假，以此循环判别完 C 类风格图片，判别完成后就对应 C 个输出通道，该模型可以同时实现多个风格域之间的字体风格转换。局部判别器则是关注于局部纹理细节的提升，使得模型能够生成更好的纹理细节。局部判别器主要负责评估生成图像的局部纹理和细节，帮助生成器更好地捕捉目标字体的艺术风格和特征。具体来说，局部判别器能够帮助生成器注意到并学习目标字体图像中的局部特征和纹理细节。通过使用局部判别器，生成器可以被引导去关注每个局部区域的特征，而不是简单地整体化地进行风格迁移，从而提高了生成图像的真实感和逼真度。风格编码器 E_S 由特定的卷积层组成，在训练过程中，风格编码器 E_S 主要负责学习参考字体（Refs）的风格特征，即从输入的 N 个参考样本中学习风格码的潜在表示。本文将风格编码器 E_S 学习到的这种潜在表示定义为风格码 Z，且输入的 N 个参考样本属于同一风格域。生成器 G 的结构类似于一个自动编码器结构的瓶颈网络，其输入是来自源字体 X_S，输出则是以风格码 Z 为条件的迁移字体 $\tilde{Y}_{X_S}$，$\tilde{Y}_{X_S}$ 与源字体 X_S 共享相同的结构，但其字体和风格与参考字体保持一致。模型框架细节如图 3 所示。

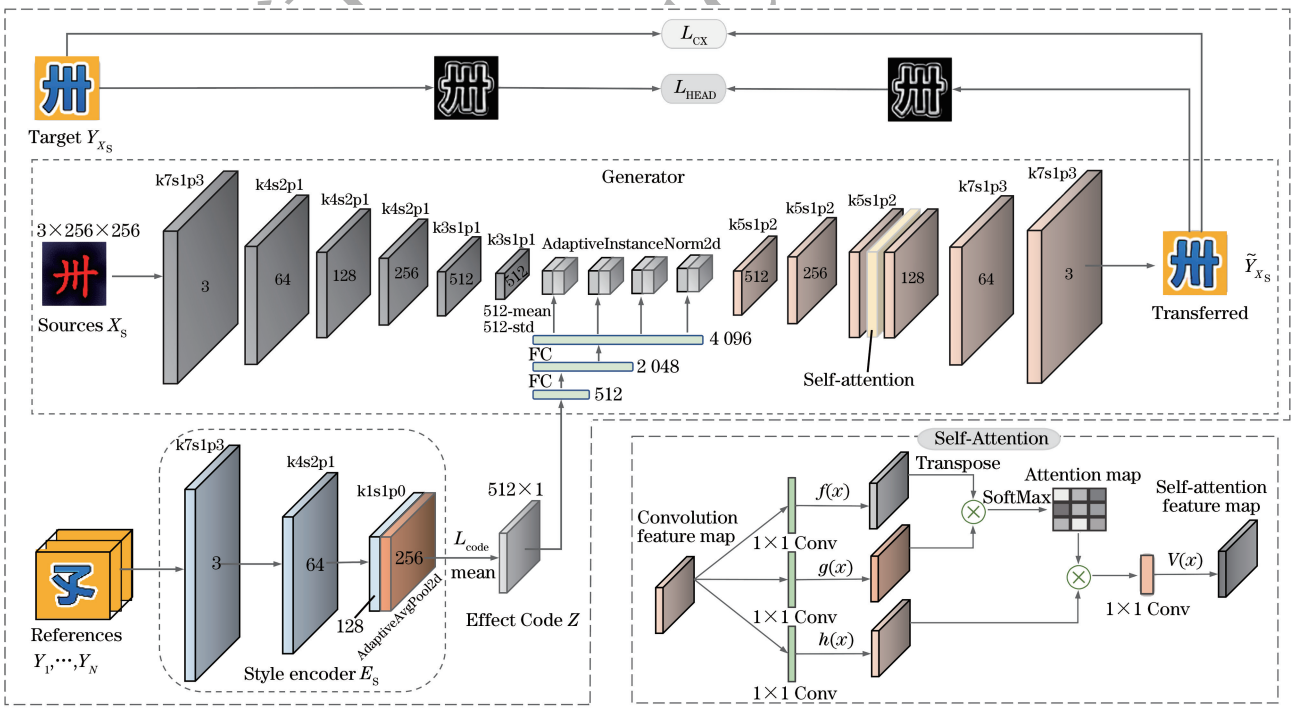


图 3 模型整体框架细节示意图

Fig. 3 Schematic diagram of the overall framework of the model

传统的 GAN 对于含有较少结构约束的类别，比如海洋、天空等，得到的结果较好；而对于含有较多几何或结构约束的类别则容易失败，比如在合成一些结构较复杂的字体时字体会变形失真。这是因为复杂的几何轮廓需要长距离依赖，卷积神经网络因为感受野大小的限制很难提取到图片中的这些长距离依赖。先前大多数的工作要么通过加深网络，要么通过扩大卷积核的尺寸来进一步提取长距离依赖，但是这会使卷积网络丧失其参数和计算的效率优势。本文在生成器的上采样部分以及鉴别器的最后一个下采样层中引入了由文献[22]提出的自注意力网络，把自注意力机制引入 GAN 的框架中，对卷

积神经网络的局限性进行补充，这有助于对图像区域中长距离和多层次的依赖关系进行建模，不仅解决了卷积结构带来的感受野大小的限制，也使得网络在生成图片的过程中能够自己学习应该关注的不同区域。如图 3 所示，自注意力机制的主要思想可以理解为：对于在网络层得到的某一个卷积特征图，将其通过 3 条支路。3 条支路中的 $f(x)$、$g(x)$ 和 $h(x)$ 是普通的 1×1 卷积块，它们的不同仅仅只是输出通道的大小而已，并最终得到带有注意力权重的特征图。在卷积所得到的特征图中，自注意力网络实质上是通过计算并预测出某个位置的像素点与其他任何位置的像素点之间的协方差矩阵，从而在模型训练的过程中达到对任意特征像素级预估的总体参考，这样使得整个网络可以学习到图像更深层级的语义信息。

对于风格图像和内容图像之间轮廓信息的提取而言，传统的大多数工作通过计算生成图像与原始内容图像之间的边缘信息差异来约束图像的边缘轮廓，使得生成图像能更好地保留原始内容图像的边缘信息。但是边缘损失函数对噪声和图像质量的敏感度较高。在噪声较多或图像质量较差的情况下，边缘损失函数可能无法有效地指导模型学习到准确的边缘信息，导致训练不稳定或者结果不理想。本文引入了 HED 模型来对风格图片和内容图片提取轮廓信息。因为在整个轮廓提取过程中，字形的背景信息会被全部移除，所以它对图片中存在的噪声有很强的抗干扰性。本文方法能够克服以上提到的稳定性不够和结果不理想的局限性，用它提取生成图像和目标图像的边缘后用于损失函数的计算，以此实现字形轮廓的约束。有关损失函数的设计在后续内容中详细介绍。

为了能够实现多个文本域之间的风格迁移，本文对文献[23]提出的基于补丁的鉴别器进行了改进，将其由单类别模式推广到多类别版本。如图 4 所示，本文将输出通道数转换为训练集的类别数，由此可以将该多类别鉴别器当作 F 个鉴别器来使用，F 即为训练集中的类别数。如图 5 所示，第 2 个鉴别器是局部鉴别器 D_{local}，主要用于优化局部的纹理细节特征。接下来将详细探讨关于目标函数的设计。

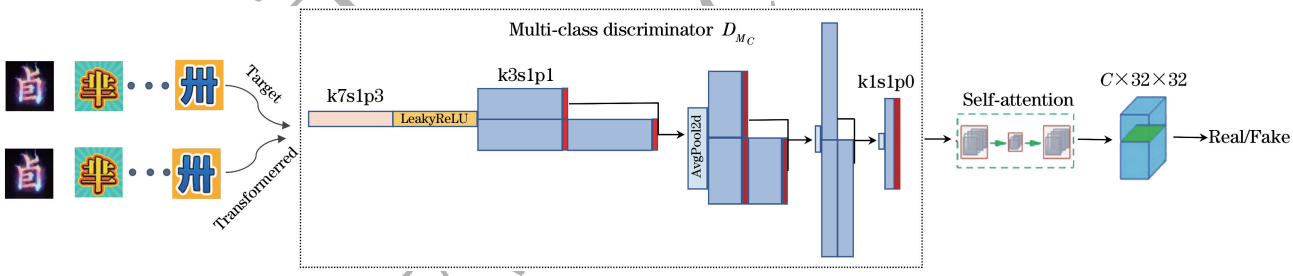


图 4 多分类判别器 D_{M_C} 的框架图

Fig. 4 Frame diagram of the multi-class discriminator D_{M_C}

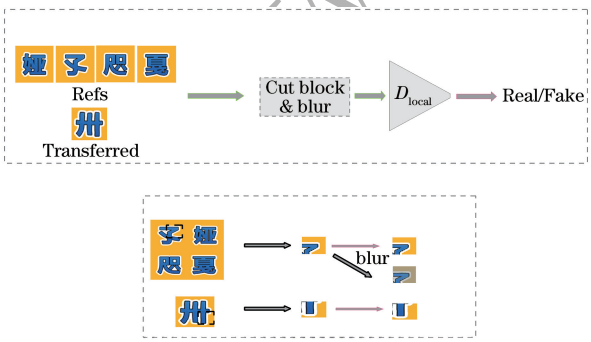


图 5 局部判别器 D_{local} 的框架图

Fig. 5 Frame diagram of the multi-class discriminator D_{local}

2.2 目标函数设计

1）风格码一致性损失 L_{code}。

编码器 E_S 提取输入参考图像的风格码。由于参考图像均属于同一文本效果域，因此编码器提取出的风格码应该具有一致性。在此前提下，本文为风格编码器设计了以下损失函数：

$$L_{\text{code}}=E_{Y_1,\cdots,Y_N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|E_S(Y_i-Y_j)\right\|_1 \quad (1)$$

式中：$E_{Y_1\cdots Y_N}$ 代表输入的参考图像。该损失函数的设计是借鉴于文献[24]的思想，该方法通过消除相似样本的特征表示来解开生成因素的纠缠。

2）字形约束损失 L_{HED}。

本文引入了 HED 模型来提取生成字体和目标字体的轮廓，并计算相应的 L_1 损失，主要目的是为了更好地约束生成图像的字形结构。本文利用 HED 网络，以生成图像和目标字形图像作为输入，经过边缘检测输出对应的二进制图像。在此过程中，字形图像的背景被移除，但字形轮廓特征被有效保留。因此，在字体迁移的过程中，可以有效地避免字体图像背景对字形结构生成的影响。在获得生成图像和目标图像的二值边缘图后，对

于不同风格域的相同字符，本文设计了以下损失函数：

$$L_{\mathrm{HED}}=E_{x\sim y,y\sim x}\left\|\mathrm{HED}(y_x)-\mathrm{HED}\left(G\left(x,\sum_{i=1}^{N}E(y_i)/N\right)\right)\right\|_1 \tag{2}$$

式中：y_x 是 Y 域 x 的目标样本，它给出了训练过程中的监督信号，如果没有 y_x，则该训练过程将变为无监督学习。根据式(1)中的假设，这里取 x 个参考样本效应码的平均值作为 Y 的最终效应码。

3)上下文损失 L_{CX}。

MECHREZ 等[25]于 2018 年提出了上下文损失。当给定 2 组特征图的集合 $X=\{x_1,x_2,\cdots,x_N\}$ 和 $Y=\{y_1,y_2,\cdots,y_N\}$，对于 Y 域中的任意元素 y_j，都可以在 X 域中找到最相似的 x_i 与之对应。先计算两者之间的余弦距离，将其归一化后转换为 2 幅图像之间的相似度。由此可以计算 2 个域中所有对应值相似度的平均值来作为 X 和 Y 之间的平均相似度，再对其取负对数就可以转换为所需的损失值。本文使用 L_{CX} 来约束生成字体 $\tilde{Y}_{X_S}$ 与目标字体 Y_{X_S} 的相似度，其计算如下所示：

$$L_{\mathrm{CX}}=\lambda_{\mathrm{CX}}\left(-\frac{1}{N}\sum_{n}\ln(r_{\mathrm{CX}}(\Psi^n(Y_{X_S}),\Psi^n(\tilde{Y}_{X_S})))\right) \tag{3}$$

式中：$\Psi^n(\cdot)$ 代表从 VGG19 的第 n 层中提取的特征；N 代表使用的卷积层的层数；λ_{CX} 表示权重。$r_{\mathrm{CX}}(X,Y)$ 表示 X 和 Y 之间的相似度。

4)局部纹理优化损失。

本文使用一个局部判别器来改善生成图像的纹理细节[22]。一方面，通过将参考字体图像剪裁成小块作为局部判别器的输入，该方法能够很好地避免由于正负样本缺乏导致训练不平衡的状况；另一方面，将生成的补丁块和经过模糊化处理的模糊块作为负样本来训练局部判别器，该方法可以使模型更好地学习到图像的局部纹理特征，有利于生成更好的纹理细节。局部纹理优化损失函数定义如下：

$$\begin{aligned}L_{\mathrm{local}}=&E_{S_{\mathrm{ref}}}[\ln(D_{\mathrm{local}}(S_{\mathrm{ref}}))]+\\&E_{\mathrm{blurl}}[\ln(1-D_{\mathrm{local}}(S_{\mathrm{blure}}))]+\\&\lambda_{\mathrm{local}}E_{S_y}[\ln(1-D_{\mathrm{local}}(S_y))]\end{aligned} \tag{4}$$

式中：S_{ref} 和 S_y 分别表示由参考字形图像和生成字形图像剪裁成的补丁块；S_{blur} 表示由 S_{ref} 进行高斯模糊后得到的模糊块；λ_{local} 表示权重。

5)对抗性损失。

为了使生成的字形图像可以无限接近真实的目标图像，本文使用了文献[26]提出的铰链版本的对抗性损失，以此来混淆鉴别器。除此之外，为了进一步确保模型训练过程的收敛性和稳定性，本文对输入的真实样本使用了文献[27]提出的梯度惩罚正则化，其主要作用是避免模型陷入局部最优而导致崩溃，即实现结构风险最小化，从而保证网络训练的稳定性。由输入的源字体图像和参考字体图像重构输出 $G\left(x,\sum_{i=1}^{N}E(y_i)\right)$，仅仅使用多类鉴别器 D_{M_C} 在通道 C(定义为 D_C)中的相应预测得分来计算其对抗损失，其损失函数表示如下：

$$\begin{aligned}L_{\mathrm{adv}}=&E_{Y_1,\cdots,Y_N}[\log_a(1-D_C(Y_i))]+\\&E_{Y_1,\cdots,Y_N}\left[D_C\left(G\left(x,\sum_{i=1}^{N}E_S\left(Y_i/N\right)\right)\right)\right]+\\&\lambda_{\mathrm{GP}}E_{Y_1,\cdots,Y_N}\|\nabla D(Y_i)\|^2\end{aligned} \tag{5}$$

式中：λ_{GP} 表示梯度惩罚正则项的权重，此处取值 $\lambda_{\mathrm{GP}}=15$；$\nabla D(Y_i)$ 表示输入参考字体图像的梯度值。

综上所述，该模型总的目标函数包括风格码一致性损失、字形约束损失、上下文损失、局部纹理优化损失和对抗性损失。总损失表示如下：

$$\begin{aligned}L(\mathrm{GAN})=&\lambda_{\mathrm{code}}L_{\mathrm{code}}+\lambda_{\mathrm{HED}}L_{\mathrm{HED}}+\lambda_{\mathrm{CX}}L_{\mathrm{CX}}+\\&\lambda_{\mathrm{local}}L_{\mathrm{local}}+\lambda_{\mathrm{adv}}L_{\mathrm{adv}}\end{aligned} \tag{6}$$

式中：λ_{code}、λ_{HED}、λ_{CX}、λ_{local} 和 λ_{adv} 表示损失函数的超参数。因此，模型可以通过以下损失函数来进行优化：

$$\min_{G,E}\max_{D_{M_C}/D_{\mathrm{local}}}L(\mathrm{GAN}) \tag{7}$$

3 实验验证

3.1 数据集

本文使用了 FET-GAN 任务中所收集的字形图像数据集。该数据集是在 TET-GAN 工作的基础上将原来配对的字形图像进行了分离，最终得到了 60 种带有不同艺术风格的字形图像以及 80 种不同字体的纯文本字形图像。整个数据集共有 108 500 张图像，每张图像的尺寸大小均被处理为 3×256×256。每种字体共包含 837 个字符，其中含有 775 个汉字、52 个大小写英文字母以及 10 个阿拉伯数字。

3.2 实验设置

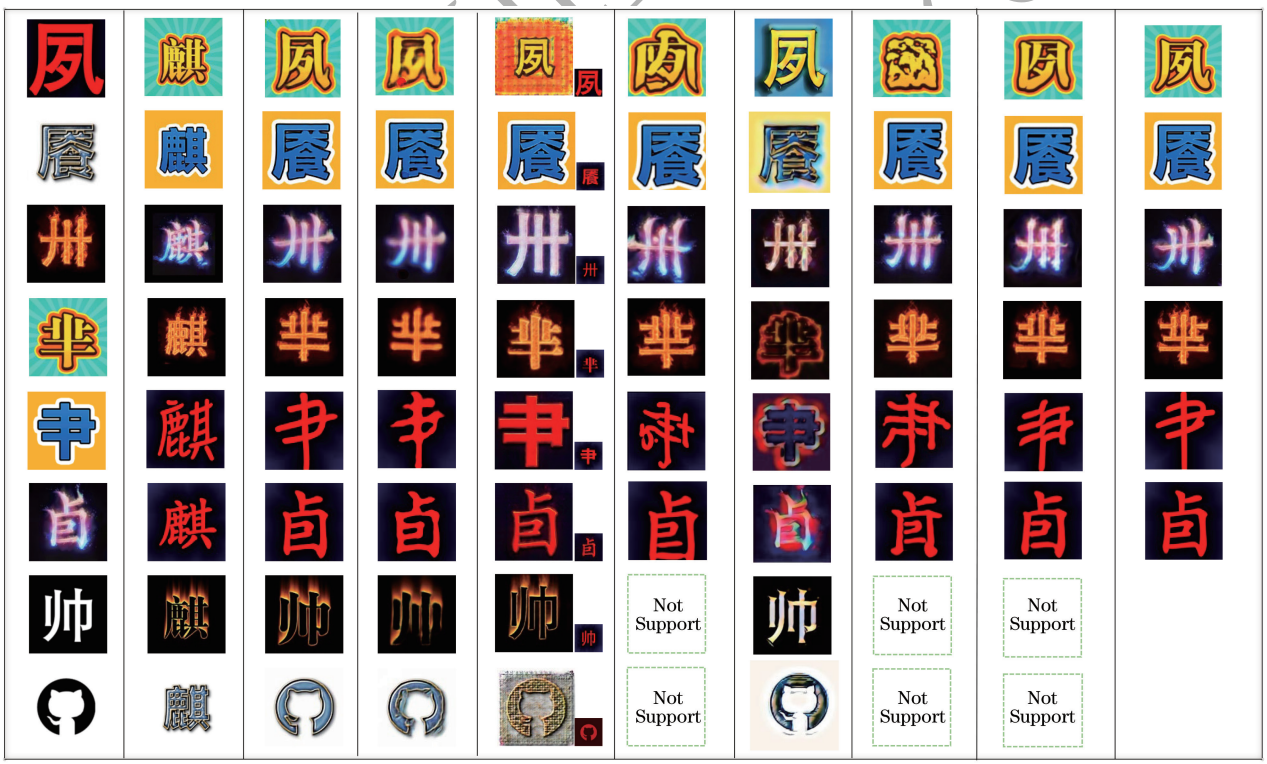
本实验基于 PyTorch 1.2，并在 V100s GPU(显存为 32 GB)上进行训练。训练过程设置输入

计算机工程 www.ecice06.com

的批量大小为 15，每次输入 4 张参考图像，整个实验的训练周期为 30。在训练时候保证一个 mini-batch 只有 Real 样本或是 Fake 样本，不把 Real 样本和 Fake 样本混合起来训练。图像首先经过预处理，将输入图像在 3 个颜色通道上进行标准化，之后将结果变换为网络能够接受的输入格式。在训练过程中通过前向传播计算风格迁移的损失函数，之后通过反向传播迭代模型参数；而后处理则是将网络输出的图像的像素值还原到标准化之前的值。

3.3 对比实验

为了更好地说明本文方法的可靠性和有效性，将本文方法与 6 种方法进行了对比，分别是：FET-GAN[23]，TET-GAN[9]，Cycle-GAN[28]，AdaIN[12]，StarGANv2[29]，U-GAT-IT[30]。实验结果如图 6 所示，其中，Ground truth 表示真实目标字体。从对比中可以看到：FET-GAN 能够保留风格图像的颜色特征，但是生成的字体产生了伪影失真；TET-GAN 虽然克服了伪影失真的局限性，但是不能很好地学习到风格图像的字体字型信息；AdaIN 不仅不能很好地学习到风格图像中的内容，而且产生的字体也存在失真和伪影；FET-GAN、StarGANv2 和 U-GAT-IT 虽然能够学到风格图像的颜色信息，但是结果出现了严重的失真，部分字体字型产生了严重的变形。从对比的结果可以看出，本文方法无论是在对图像风格的传递方面还是对字体的变换方面均有着一定的优势，尤其是在从纯文本字形图像到风格字形图像之间的传递方面优势最为显著。



(a)源字体 (b)参考字体 (c)本文方法 (d)FET-GAN (e)TET-GAN (f)Cycle-GAN (g)AdaIN (h)StarGANv2 (i)U-GAT-IT (j)Ground truth

图 6 本文方法与其他 6 种方法的实验结果比较

Fig. 6 Comparison of experimental results of the proposed method and other six methods

3.4 评价指标

本文采用了字体迁移任务中常用的 3 种评价指标：结构相似度（SSIM）[31]，峰值信噪比（PSNR）[31]，用于评估生成图像质量的度量标准 FID（Fréchet Inception Distance）[32]。其中：SSIM 的值越大，表示输出图像和目标图像的结构差异越小，即图像质量越好；PSNR 的值越大，说明生成图像的失真越少[33]；FID 的值越小说明图片的质量和多样性越好。实验结果的评价指标如表 1 所示，其中，加粗数据表示最优值。

表 1 本文模型和其他 6 种方法的评价指标比较

Table 1 Comparison of metrics of the proposed model and other six methods

Model	SSIM	PSNR/dB	FID
AdaIN	0.229	10.343	194.652
Cycle-GAN	0.329	10.775	96.146
TET-GAN	0.441	13.976	192.428
StarGANv2	0.455	14.012	131.645
U-GAT-IT	0.557	15.228	91.435
FET-GAN	0.628	17.063	171.762
Ours	**0.745**	**18.668**	**72.355**

3.5 多域转换

本文提出的模型可实现多个风格域之间的风格迁移，这得益于多分类鉴别器 D_{M_C} 。如图 4 所示，D_{M_C} 的输出通道数设定为训练集的类别数 C，因此可以把这样的多分类鉴别器看作是除最后一个卷积层之外其余层共享参数的 C 个鉴别器。C 是训练集的类别数，每个类别属于不同的字体域。图 7 中可以非常直观地看出每一种字体转换为其他 8 种风格时的结果。除了汉字字符，本文还进行了英文字符和阿拉伯数字的风格迁移实验，如图 8 所示，每一种风格的合成字符都显示在第 2 行，第 1 行显示的是 Ground truth。

图 7 多种风格域之间的相互转换

Fig. 7 Interconversion between various style domains

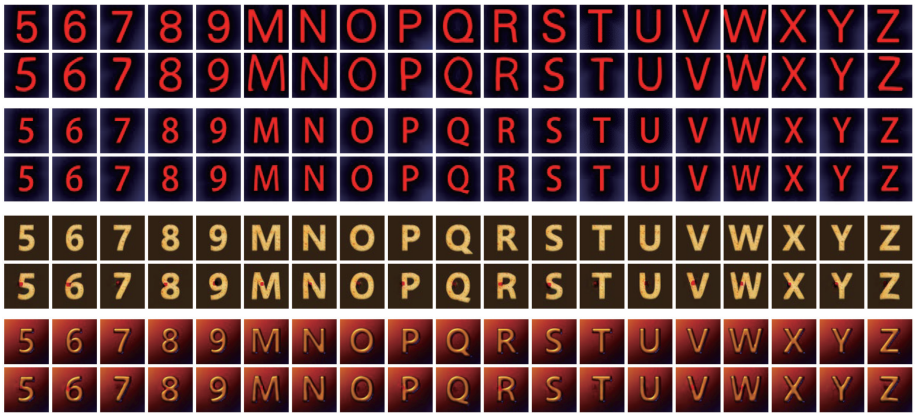


图 8 本文模型在英文和阿拉伯数字字符上的迁移效果

Fig. 8 The transfer effect of the proposed model on English and Arabic numerals

3.6 消融研究

3.6.1 消融实验

消融实验结果如图 9 所示，其中：w/o 表示去除；第 1 行为真实目标字体；第 2 行为本文方法结果；与第 2 行相比，第 3 行中去除了自注意力和光谱归一化，由于 GAN 训练的不稳定，迁移的结果出现了黑斑或白班，并且部分图像的亮度变暗；与第 3 行相比，第 4 行中去除了自注意力和局部纹理优化损失，迁移的结果出现了颜色分布不均匀，生成的纯文本字形图像中字形笔画结构出现了缺失；第 5 行为基线模型的结果，与第 5 行相比，其去除了字形约束损失，迁移的结果中出现了模糊和伪影，而且部分结果中字形结构严重缺失。

图 9 本文模型消融实验结果

Fig. 9 Ablation experiment results of the proposed model

3.6.2 消融实验定量分析

对消融实验结果进行评估的量化结果如表 2 所示，其中，w/o 表示去除。表中的指标直观地反映了每个组件对整个模型的影响，从表 2 中的数据可以清晰地看出：字形约束损失能够使生成的字形与目标字形尽可能相似；局部纹理优化损失和上下文损失对提升生成图像的局部纹理细节质量也有很大的贡献；自注意力的引入有助于提升生成图像的质量。

表 2 3 组消融实验的定量结果

Table 2 Quantitative results of three groups of ablation experiments

Model	SSIM	PSNR/dB	FID
w/o Self-attention, L_{local}, L_{CX}, L_{HED}	0.657	16.358	159.154
w/o Self-attention, L_{local}	0.683	17.564	112.963
w/o Self-attention	0.704	18.126	90.877
Full Net	**0.745**	**18.668**	**72.355**

3.7 参数分析

本文算法的核心为字形约束模块，其对字体迁移结果的好坏起着至关重要的作用。为了进一步探讨参数 λ_{CX} 对字形结构的影响，在式(6)中，在保持参数 λ_{HED} 不变的前提下设置多组不同的超参数 λ_{CX} 值进行参数分析实验，实验的定量和定性结果如表 3 和图 10 所示。可以看出，随着 λ_{CX} 值的增加，生成图像的字形结构越来越完整，但是当超过一定值的时候，生成图像的视觉效果会逐渐降低，当 $\lambda_{CX}=0.15$ 时，L_{CX} 的效果最好。

表 3 不同 λ_{CX} 值的定量比较

Table 3 Quantitative comparison of different λ_{CX} values

Metrics	SSIM	PSNR/dB	FID
$\lambda_{CX}=0.05$, $\lambda_{HED}=1$	0.450	12.731	190.49
$\lambda_{CX}=0.10$, $\lambda_{HED}=1$	0.619	15.355	142.72
$\lambda_{CX}=0.15$, $\lambda_{HED}=1$	**0.656**	**16.271**	**129.574**
$\lambda_{CX}=0.20$, $\lambda_{HED}=1$	0.628	15.225	171.854
$\lambda_{CX}=0.25$, $\lambda_{HED}=1$	0.603	15.55	196.687

同理，为了进一步研究字形约束损失的影响，本文进行了另一组参数分析实验：在式(6)中，保持 λ_{CX} 的值不变，通过设置多组不同的 λ_{HED} 值进行实验。此处，设置 λ_{CX} 和 λ_{HED} 的不同组合是为了在保持字形图像结构的完整性和字形易读性之间找到某种平衡。实验的定性结果展示如图 11 所示，用于测量生成图像质量的 3 个评价指标如表 4 所示，所有实验均表明，当 $\lambda_{CX}=0.15$ 且 $\lambda_{HED}=2$ 时，字体风格迁移的效果最佳。

计算机工程 www.ecice

图 10 超参数 λ_{CX} 分析的结果比较

Fig. 10 Results comparison of hyperparameter λ_{CX} analysis

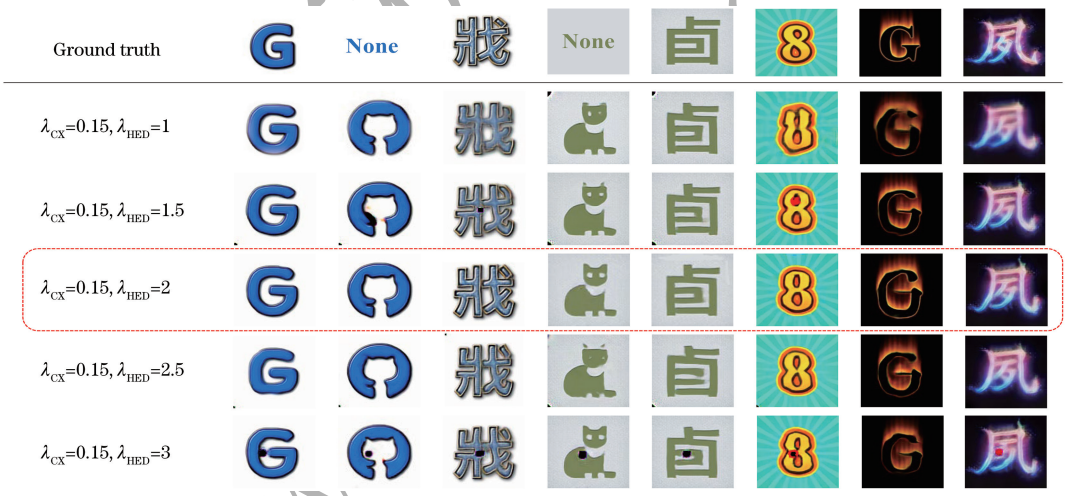


图 11 超参数 λ_{HED} 分析的结果比较

Fig. 11 Results comparison of hyperparameter λ_{HED} analysis

表 4 不同 λ_{HED} 值的定量比较

Table 4 Quantitative comparison of different λ_{HED} values

Metrics	SSIM	PSNR/dB	FID
$\lambda_{CX}=0.15, \lambda_{HED}=1.0$	0.626	16.575	136.965
$\lambda_{CX}=0.15, \lambda_{HED}=1.5$	0.645	16.611	149.067
$\lambda_{CX}=0.15, \lambda_{HED}=2.0$	**0.686**	**16.757**	**121.765**
$\lambda_{CX}=0.15, \lambda_{HED}=2.5$	0.672	16.621	156.062
$\lambda_{CX}=0.15, \lambda_{HED}=3.0$	0.636	16.026	185.129

3.8 模型应用:自动补充和完善字体库

为了推广模型的泛化能力和应用场景,本文进行了一个纯文本字体之间的迁移实验。基于 80 种字体数据集的实验结果如图 12 所示,其中,第 1 行和第 1 列分别表示参考字体和源字体,默认共进行了 30 个周期的迭代。图 12 中的结果表明,本文模型在纯文本字体迁移任务中依然能得到较好的结果。这有助于帮助自动补充和完善字体库,可以极大地减轻字体设计成本。

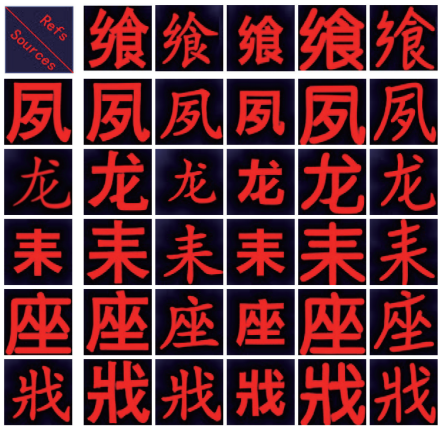


图 12 本文模型推广到自动补充字体库的示例

Fig. 12 Example of extending the proposed model to auto-complement font libraries

4 结束语

本文在现有研究方法的基础上，提出了基于字形约束和注意力的艺术字体风格迁移方法，构建了新颖有效的网络框架。该模型能够在多个文本效果域之间实现带有字体变换的字形图像风格迁移，并且在文字风格迁移任务中，和其他对比方法相比均具有一定的优势，迁移结果能够保持很好的字形结构。本文方法 FID 值为 72.355，低于对比实验中最好的结果 91.435。此外，本文模型可以推广到一些新的字符迁移任务中，在生成新字体和补充字体库方面具有重要的意义。应当指出的是，如何仅利用很少的样本就可以实现大量艺术字形矢量图像的自动生成，这是未来将要探索的一个颇具挑战性的方向。

GAN 虽然是一种强大的生成模型，但是在训练过程中，生成器和判别器之间的动态平衡可能会出现波动，导致模型性能不稳定，而且网络架构和超参数对于 GAN 的成功至关重要，但这通常需要大量的实验和调整，本文实验也存在以上提到的问题。在下一步工作中，将把生成器改为训练过程更为可控的扩散模型。

参考文献

[1] 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. Washington D. C., USA: IEEE Press, 2016: 2414-2423.

[2] GATYS L A, ECKER A S, BETHGE M. Texture synthesis using convolutional neural networks[EB/OL]. [2023-05-20]. https://arxiv.org/abs/1505.07376.

[3] LI Y, WANG N, LIU J, et al. Demystifying neural style transfer[EB/OL]. [2023-05-20]. https://arxiv.org/abs/701.01036.

[4] LI Y, FANG C, YANG J, et al. Universal style transfer via feature transforms[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2017: 385-395.

[5] CAMPBELL N D F, KAUTZ J. Learning a manifold of fonts[J]. ACM Transactions on Graphics, 2014, 33(4): 1-11.

[6] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2014: 2672-2680.

[7] YANG S, LIU J Y, WANG W J, et al. TET-GAN: text effects transfer via stylization and destylization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 1238-1245.

[8] LI C, WAND M. Precomputed real-time texture synthesis with Markovian generative adversarial networks[C]//Proceedings of ECCV 2016. Berlin, Germany: Springer, 2016: 702-716.

[9] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C]//Proceedings of the IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 1501-1510.

[10] LIU M Y, HUANG X, MALLYA A, et al. Few-shot unsupervised image-to-image translation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2019: 10551-10560.

[11] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 4401-4410.

[12] WANG Z, ZHAO L, CHEN H, et al. Diversified patch-based style transfer with shifted style normalization[EB/OL]. [2023-05-20]. https://arxiv.org/abs/2101.06381v1.

[13] KALISCHEK N, WEGNER J D, SCHINDLER K. In the light of feature distributions: moment matching for neural style transfer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2021: 9382-9391.

[14] CANHAM T D, MARTÍN A, BERTALMÍO M, et al. Using decoupled features for photo-realistic style transfer[EB/OL]. [2023-05-20]. https://arxiv.org/abs/2212.02953.

[15] HONG K, JEON S, LEE J, et al. AesPA-net: aesthetic pattern-aware style transfer networks[EB/OL]. [2023-05-20]. https://arxiv.org/abs/2307.09724.

[16] 让孝迪. 基于生成对抗网络的无监督艺术图像风格迁移[D]. 烟台：烟台大学，2023.
RANG X D. Unsupervised art image style transfer based on generative confrontation network[D]. Yantai: Yantai University, 2023. (in Chinese)

[17] 过劲. 基于生成对抗网络的艺术风格图像迁移研究[D]. 南昌：南昌大学，2023.
GUO J. Research on image migration of artistic style based on generative confrontation network[D]. Nanchang: Nanchang University, 2023. (in Chinese)

[18] TOGO R, KOTERA M, OGAWA T, et al. Text-guided style transfer-based image manipulation using multimodal generative models[J]. IEEE Access, 2021, 9: 64860-64870.

[19] CHEN H B, ZHAO L, ZHANG H M, et al. Diverse image style transfer via invertible cross-space mapping[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2021: 14860-14869.

[20] CAO J. Hierarchical-based calligraphy style transfer[J]. World Scientific Research Journal, 2021, 7(5): 430-439.

[21] LI W, HE Y X, QI Y W, et al. FET-GAN: font and effect transfer via K-shot adaptive instance normalization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(2): 1717-1724.

[22] ZHANG H, GOODFELLOW I, METAXAS D, et al. Self-attention generative adversarial networks[C]//Proceedings of International Conference on Machine Learning. [S. l.]: PMLR, 2019: 7354-7363.

[23] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 1125-1134.

[24] KULKARNI T D, WHITNEY W, KOHLI P, et al. Deep convolutional inverse graphics network[EB/OL]. [2023-05-20]. https://arxiv.org/abs/1503.03167.

[25] MECHREZ R, TALMI I, ZELNIK-MANOR L. The contextual loss for image transformation with non-aligned data[C]//Proceedings of the European Conference on

计算机工程 www.ecice06.com

Computer Vision. Berlin, Germany: Springer, 2018: 768-783.
[26] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral normalization for generative adversarial networks[EB/OL]. [2023-05-20]. https://arxiv.org/abs/1802.05957.
[27] MESCHEDER L, GEIGER A, NOWOZIN S. Which training methods for GANs do actually converge[C]// Proceedings of International Conference on Machine Learning. [S. l.]: PMLR, 2018: 3481-3490.
[28] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 2223-2232.
[29] CHOI Y, UH Y, YOO J, et al. StarGAN v2: diverse image synthesis for multiple domains[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 8188-8197.
[30] KIM J, KIM M, KANG H, et al. U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation[EB/OL]. [2023-05-20]. https://arxiv.org/abs/1907.10830.
[31] HORE A, ZIOU D. Image quality metrics: PSNR vs. SSIM[C]//Proceedings of the 20th International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2010: 2366-2369.
[32] PREUER K, RENZ P, UNTERTHINER T, et al. Fréchet ChemNet Distance: a metric for generative models for molecules in drug discovery[J]. Journal of Chemical Information and Modeling, 2018, 58(9): 1736-1741.
[33] 钱旭淼, 段锦, 刘举, 等. 基于注意力特征融合的图像去雾算法[J]. 吉林大学学报(理学版), 2023, 61(3): 567-576.
QIAN X M, DUAN J, LIU J, et al. Image dehazing algorithm based on attention feature fusion[J]. Journal of Jilin University(Science Edition), 2023, 61(3): 567-576. (in Chinese)

编辑 金胡考

计算机工程 www.ecice06.com