DOI:10.19651/j.cnki.emt.2415596

基于多层次通道注意力网络的少样本字体生成*

邱燕波[1] 储开斌[1] 张 继[2] 冯成涛[1]
(1.常州大学微电子与控制工程学院 常州 213159; 2.常州大学计算机科学与人工智能学院 常州 213159)

摘 要: 为提升字体生成的图像质量,减少字体设计的人工成本,提出基于多层次通道注意力网络的少样本字体生成的方法。首先,该方法通过风格感知注意力模块获取重要的局部特征;然后设计了一个多层次的注意力机制,较浅的层只能观察图像的局部特征,而较深的层可以观察到图像的全部特征,通过聚合不同层次的局部特征来构建新的风格特征。最后,使用了内容损失函数、风格损失函数和 L_1 损失函数优化模型的参数,稳定网络的训练,使生成图像在内容和风格上与目标字体保持一致。实验结果表明,该方法在未知样式的字体和未知内容的字体具有很强的泛化性。相比于其他方法,所提出的方法表现出更好的实验结果,能保持内容结构的完整和字形风格的准确性。

关键词: 字体生成;生成对抗网络;风格迁移
中图分类号: TP391 **文献标识码**: A **国家标准学科分类代码**: 510.4050

Few-shot font generation for multilevel channel attention networks

Qiu Yanbo[1] Chu Kaibin[1] Zhang Ji[2] Feng Chengtao[1]
(1. School of Microelectronics and Control Engineering, Changzhou University, Changzhou 213159, China;
2. School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213159, China)

Abstract: In order to improve the image quality of font generation and reduce the labour cost of font design, a method for few-shot font generation based on multilevel channel attention network is proposed. Firstly, the method acquires important local features through the style-aware attention module; then a multilevel attention mechanism is designed, where shallower layers can only observe the local features of the image, while deeper layers can observe all the features of the image, and new stylistic features are constructed by aggregating the local features of different levels. Finally, a content loss function, a style loss function and a L_1 loss function are used to optimise the parameters of the model and stabilise the training of the network so that the generated images are consistent with the target font in terms of content and style. The experimental results show that the method has a strong generalisation to fonts of unknown style and fonts of unknown content. Compared to other methods, the proposed method shows better experimental results that maintain the integrity of the content structure and the accuracy of the font style.

Keywords: font generation;generative adversarial nets;style transfer

0 引 言

作为书写和表达的基本元素,字体不仅仅是文字的外在形态,更是文化、历史和品牌形象的重要载体。字体生成在古籍古迹字体补全、个性化字体设计、品牌识别和文化艺术交流等方面有着广泛的应用。随着科技和经济的迅猛发展,文化、传媒、商业、广告等各行业对汉字的需求急剧增加。而设计一套完整的字库需要花费大量的人力物力,这与现有的字体生成技术形成强烈的供需矛盾。因此,设计一种低成本、高效率和高质量的字体生成方法尤为重要。

随着计算机视觉[1-2]得到高速发展,字体风格迁移成为一门涵盖计算机科学、艺术设计和人机交互领域的新兴技术。在字体生成背后,深度学习[3-4]发挥着关键的作用。目前主流的方法是使用深度学习来实现字体生成任务,这主要分为有监督学习和无监督学习两种方式。

在无监督学习方式中,Zhu 等[5]提出一种循环一致对抗网络(cycle-consistent adversarial networks, CycleGAN),该方法使用了两个域来实现图像的风格转换,但这种方法只能

收稿日期:2024-03-07
*基金项目:江苏省基金项目(2019JSJG243)、江苏省高等学校自然科学研究面上项目(19KJB510017)、江苏省研究生科研与实践创新计划项目(KYCX23_3182,YPC23020168)资助

获得编码器提取的低维空间特征，生成的图像存在伪影、笔画错乱等情况。Chang 等[6]在 CycleGAN 的传输块中使用稠密网络[7]（densely connected convolutional networks，DenseNet）特征重复利用，在一定程度上改善了生成器的生成效果，但生成的风格化图像中仍然存在源域的风格，这说明该方法并没有完全实现风格迁移。可变形生成网络（deformable generative networks for unsupervised font generation，DG-Font）[8]分离内容特征和风格特征，然后使用特征变形跳跃连接对内容图像的底层特征进行变换，最后使用自适应实例规范化[9]（adaptive instance normalization，AdaIN）将风格特征和征与内容特征相结合。刘等[10]在 DG-Font 的框架上使用风格注入网络和变形注意力跳跃连接模块。曾等[11]提出了一种基于生成对抗网络的无监督字体生成方法，该方法提取内容图像的内容特征和目标字体的风格特征，然后使用风格注意网络[12]（style-attentional networks，SANet）将内容特征与风格特征融合在一起，同时使用相对判别器稳定模型的训练。这些方法都是将内容特征和风格特征简单的拼接在一起，在捕获风格特征时没有充分利用参考风格集的共同特征，削弱了生成图像风格特征信息，使生成图像与目标字体图像风格不一致。

在有监督学习方式中，Zi2zi 项目没有以论文的形式发表，但其带动了字体生成在图像翻译中的应用，它在 pix2pix[13]的基础上加入了一种风格类别标签，并通过配对数据集的有监督方式训练网络来实现一对多的字体风格迁移。这种方法虽然实现了一对多的字体生成，但在训练网络时需要花费非常多的时间，最后这种方法只能从特定的字体样式中获得良好的效果，如果字体结构过于复杂，则生成的图像质量较差。陈杰夫等[14]在多领域图像到图像翻译（unified generative adversarial networks for multi-domain image-to-image translation，StarGAN）[15]中加入一种样式指定机制实现风格迁移，它与 zi2zi 一样都可以进行一对多的风格迁移，但图像质量并不理想。随着字体风格复杂度的提升和对生成图像质量要求的提高，这些模型需要更多不同风格的样本数据来训练模型学习目标风格的结构空间分布，而收集和标记大量的训练样本需要耗费大量的人力物力。王等[16]在生成器中使用了笔画先验信息引导模型生成与目标字体一致的字体，同时使用融合注意力机制的跳跃连接模块来解决字体不完整和模糊的问题，但是笔画先验信息的制作是非常昂贵的。

以 CycleGAN 为主的无监督字体生成方法生成的字体会遗留一些内容特征。因此，本文使用少量配对样本的有监督学习方式来实现字体生成，大大减少了数据集的数量和模型的预处理要求。现有的少样本字体生成方法主要存在两个问题：一是无法正确衡量字体风格的决定因素，即无法判断需要转换的字体风格应该关注全局特征还是局部特征；二是不能充分学习风格参考集中的共同特征，主要原因是这些方法认为风格图像中的每个区域所作出的贡献是相同的，容易把风格参考集中的每个文本图像都赋予相同的权重来构建新的风格特征。

为了解决这些问题，本文提出一种基于多层次通道注意力网络的少样本字体生成方法，该方法主要是对跨语言字体样式转移（cross-language font style transfer，FTransGAN）[17]进行改进的。FTransGAN 是第一个实现跨语言字体生成的模型，它使用上下文感知模块和层注意力模块来捕获全局和局部特征。Yu 等[18]提出了一种用于文本分类的上下文感知的视觉注意力网络，该网络能根据上下文语境来区分重要的单词和语句。针对第一个问题，本文提出了一个风格感知注意力模块，该模块修改了上下文感知的视觉注意力模块[18]的输入和输出，并引入了一个风格向量来衡量每个区域的贡献，使网络能获取重要的局部特征，增强了网络提取特征的能力；同时保留了 FTransGAN 多层次注意力形式的思想，将风格感知注意力模块进一步设计成多层次的注意力机制，通过获取不同层次的局部特征来捕获全局特征。针对第二个问题，本文舍弃 FTransGAN 中的层注意力模块，借鉴卷积注意力模块（convolutional block attention module，CBAM）[19]的思想获取每个通道的重要程度，并将这些重要程度与每个特征相乘输出通道注意力映射，它可以自适应的调整通道维度上的特征；然后将风格编码器的输出与通道注意力映射简单连接，防止梯度消失，提高模型生成质量。同时，引入了内容判别器和风格判别器，确保生成的图像在内容和风格上与目标字体保持一致，进一步提高模型生成图像质量。

1　基于多层次通道注意力网络的少样本字体生成

本文提出的基于多层次通道注意力网络的少样本字体生成模型有一个生成器 G 和两个判别器。模型的整体框架如图 1 所示。生成器采用编码-解码的形式，这有助于在上下文融合过程中捕获不同尺度的特征。生成器是由内容编码器、风格编码器、基于风格感知的多层次通道注意力网络和解码器组成。风格编码器用于学习给定风格图像的特征，内容编码器用于学习字符内容的结构信息。

字体生成主要包括 4 个阶段。第 1 阶段是提取风格图像和内容图像的特征。利用风格编码器和内容编码器分别提取风格特征和内容特征。第 2 阶段是捕获局部和全局特征。将获取的风格特征输入多层次通道注意力网络中进一步获取不同层次的风格特征和空间信息，较浅的层只能观察到图像的局部特征，而较深的层可以观察到整个图像的特征。第三阶段是目标字体生成。内容特征与捕获的局部和全局特征进行简单拼接并输入到解码器中生成目标字体图像。第四阶段是参数优化。将生成的字体图像输入到内容判别器和风格判别器，区分生成的图像在内容和风格上是否与目标字体保持一致，并使用内容损失、风格损失和 L_1 损失帮助模型实现参数优化，提高模型的生成质量。

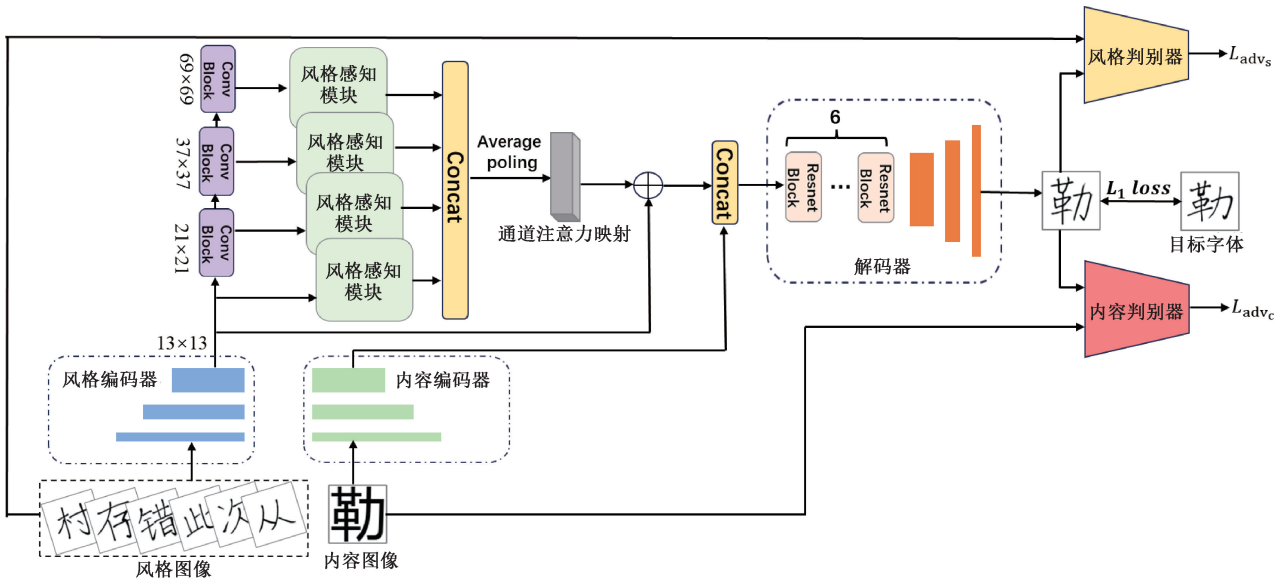

图 1 模型的整体框架

1.1 编码器和解码器

风格编码器和内容编码器具有相似的网络结构,但它们的输入不同。编码器由 3 个卷积层组成,在每个卷积层后面跟着一个批处理归一化层和 ReLU 层。

解码器由 6 个残差块和 3 个转置卷积层组成,在前两个转置卷积层后面都跟着一个批处理归一化层和 ReLU 函数,最后一个转置卷积层后面跟着 Tanh 函数。3 个转置卷积层用于对特征映射进行上采样。

1.2 风格感知注意力模块

风格感知注意力模块的网络结构如图 2 所示。输入是每个卷积块给出的特征映射,它可以保持每个区域完整的空间信息,然后将每个区域的特征向量表示为 $\{v_r, r = 1,2,\cdots,R\}$。特征映射的大小为 $C \times H \times W$,其中 C,H,W 分别表示通道数、高度和宽度。自注意力机制 f_A 是用于计算并更新特征向量 $\boldsymbol{v}_r$ 的状态,同时将相关的风格信息纳入到特征向量中。

$$\overleftrightarrow{\boldsymbol{v}}_r = f_A(\boldsymbol{v}_r) \tag{1}$$

式中:$\overleftrightarrow{\boldsymbol{v}}_r$ 包含了周围区域重要的风格信息。

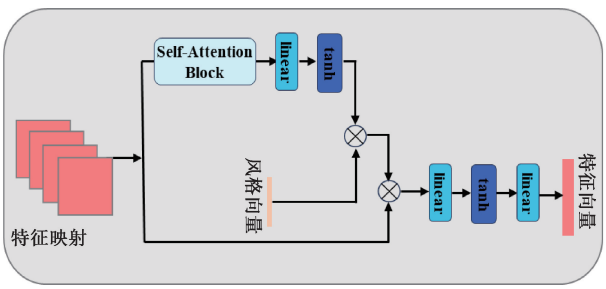

图 2 风格感知注意力模块

$$u_r = \tanh(w_a \overleftrightarrow{\boldsymbol{v}}_r + b_a) \tag{2}$$

$$q_r = soft\max(u_r^T \boldsymbol{u}_s) \tag{3}$$

$$p = \sum_{r=1}^{H \times W} q_r v_r \tag{4}$$

新特征向量 $\overleftrightarrow{\boldsymbol{v}}_r$ 输入到单层感知器中获得 u_r 作为 $\overleftrightarrow{\boldsymbol{v}}_r$ 的潜在表示,然后将 $\overleftrightarrow{\boldsymbol{v}}_r$ 的潜在表示嵌入到风格向量 $\boldsymbol{u}_s$,并为每个区域分配一个注意力分数,用于建模该区域与风格向量 $\boldsymbol{u}_s$ 之间的关系,使网络可以关注重要的区域。然后计算了一个特征映射 p 作为每个区域的加权和。

$$T = \tanh(w_c p + b_c) \tag{5}$$

$$f = T \times w_d \tag{6}$$

最后将特征映射 p 输入到多层感知器中,通过优化参数来实现 MPL 中注意力函数的自动学习,从而得到最终的风格特征。其中 w_a,w_c,w_d,b_a,b_c 均为感知器(MLP)的参数。由于网络中有 4 个并行的风格感知注意模块,最终可以获得 4 个不同层次的风格特征 f_1,f_2,f_3,f_4。

1.3 多层次的注意力机制

字体风格通常包括局部特征(如笔画、厚度、粗细等),也包括全局特征(如形状、效果等)。对于不同风格的字体,模型应该关注的是局部特征还是全局特征,这取决于风格图像本身。例如在处理手写字体时,每个人的写字方式都不同,导致手写字体存在很大的差别(例如线条粗细和厚度),此时模型会侧重于字体的局部特征;而对于印刷字体,模型则会侧重关注全局特征。

基于这一假设,本文把风格感知注意力模块设计成多层次的注意力形式,使网络同时挖掘更高层次的风格信息和空间信息。多层次的注意力机制的网络结构如图 3 所示。风格感知注意力块用于寻找与风格信息相关的重要区域,并学习这些区域的风格特征。

多层次的注意力机制是以风格感知注意力模块为基

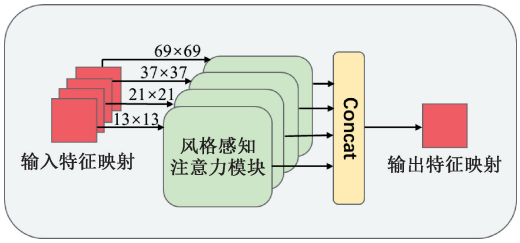

图 3　多层次的注意力机制

础，其目的是通过联合注意力学习将不同层次的风格特征合并。它有 4 个并行的风格感知注意力模块，在进行多次卷积后分别获得 13×13，21×21，37×37 和 69×69 的接受域。较浅的层只能观察到图像的局部特征，而较深的层可以观察到图像的全部特征。不同接受域的风格感知模块可以学习关注不同关系的风格特征，然后通过结合不同层次的风格特征来实现全局特征的获取，这种结合局部和全局特征的方式能更好地从图像中学习细微的风格表示。在获取 4 个不同层次的风格特征 f_1,f_2,f_3,f_4 后，将 4 个风格特征拼接成一个矩阵向量 $\boldsymbol{R}$，该过程可以定义为：

$$\boldsymbol{R}=MultiHead(f_1,f_2,f_3,f_4)=Concat(f_1,f_2,f_3,f_4) \tag{7}$$

1.4　基于风格感知的多层次注意力网络

FTransGAN 没有充分利用风格图像集的共同特征，例如将宋体“草”字转换为黑体字，其中黑体样式包含“我”、“蕊”、“中”等多个字符，“蕊”字应当是贡献最大的，因为它们具有相同的偏旁“艹”。因此，本文将 4 个特征向量拼接以后再经过平均池化和多层神经网络以及一个隐藏层来获取每个通道的重要程度，随后将这些重要程度与每个特征相乘输出通道注意力映射，它可以自适应的调整通道维度上的特征。这种类似于 CBAM 的方式可以使网络重点关注重要的风格特征。最后和风格编码器的输出简单连接。多层次通道注意力网络如图 4 所示。

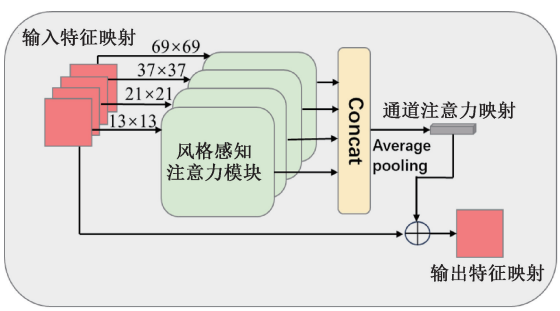

图 4　基于风格感知的多层次通道注意力网络

该过程的计算为：

$$z_m=z_s+\boldsymbol{R}\times\sigma(MLP(AvgPool(\boldsymbol{R})))=z_s+\boldsymbol{R}\times\sigma(w_1(w_0(\boldsymbol{R}^c_{avg}))) \tag{8}$$

式中：z_s 是风格编码器的输出，σ 是 sigmoid 函数，w_0 和 w_1 是 MLP 的权重，ReLU 函数跟在 w_0 后面。

1.5　损失函数

在训练过程中，模型从给定的标准字体中随机选择一个字符图像 c 输入到内容编码器中获取字符的结构信息，而风格图像 s 则是从样式字体中随机选择 6 个输入到风格编码器中获取样式特征。汉字结构复杂，仅用一个图像难以获取大量的样式特征。因此，本文将风格图像数量设置为 $k=6$。样式特征输入到基于风格感知的多层次通道注意力网络中获取每个通道的重要程度，并将这个重要程度与每个特征相乘，从而使网络重点关注一些特征通道。最后将这些特征与内容特征拼接在一起输送到解码器中生成风格化图像。

为了训练所提出的模型，使用的损失函数由 3 部分组成，分别是 L_1 损失、风格损失 L_s 和内容损失 L_c。其中风格损失 L_s 和内容损失 L_c 均使用铰链损失[20]计算，使模型更稳定地收敛和快速训练，提高生成图像的质量和模型的稳定性。

内容损失：在字体生成任务中，内容图像与生成的字体必须保持相同的字符结构。然而简单的使用像素级损失函数来约束内容特征的损失并不能产生理想的效果。因为生成的字体不只是在像素级上与输入的字体相似，它包含了一些样式细节。如果简单的使像素级损失，可能会完全忽略这些细节。因此，在训练网络的时候，内容编码器会剔除风格特征，只对内容特征进行学习。

公式如下：

$$L_{contentG}=-E_{x,c\sim p(x,c)}[D_c(x,c)] \tag{9}$$

$$L_{contentD}=-E_{\hat{x},c\sim p(\hat{x},c)}[\min(0,D_c(\hat{x},c)-1)]-E_{x,c\sim p(x,c)}[\min(0,-D_c(x,c)-1)] \tag{10}$$

$$L_{adv_c}=L_{contentD}+L_{contentG} \tag{11}$$

风格损失：与内容损失类似，通过最小化风格图像与生成图像间的风格损失来优化风格编码器，强制风格编码器保留目标字体的样式。

公式如下：

$$L_{styleG}=-E_{x,s\sim P(x,s)}[D_s(x,s)] \tag{12}$$

$$L_{styleD}=-E_{\hat{x},s\sim P(\hat{x},s)}[\min(0,D_s(\hat{x},s)-1)]-E_{x,s\sim P(x,s)}[\min(0,-D_s(x,s)-1)] \tag{13}$$

$$L_{adv_s}=L_{styleD}+L_{styleG} \tag{14}$$

L_1 损失：其次在训练过程中，还使用 L_1 损失鼓励生成器生成与真实图像相似的图像。

$$L_1=E_{x,\hat{x}\sim P(x,\hat{x})}\|x-\hat{x}\| \tag{15}$$

式中：x 是生成的图像，$\hat{x}$ 是真实的图像。

完整的目标函数如下：

$$L=\lambda_1L_1+\lambda_sL_{adv_s}+\lambda_cL_{adv_c} \tag{16}$$

其中，λ_1,λ_s 和 λ_c 是目标函数的权重，用于平衡网络的训练。

2　实验及结果分析

2.1　数据集

为了评估模型的生成能力，收集了一个包含 845 种字

体的数据集,包括手写体、印刷体以及艺术字体等,每种字体大约有993个常用汉字。通过在每个图像周围设置一个边界框并调整其大小来处理数据集,使其维度达到64像素;然后,通过填充创建64×64的字体图像。在输入到模型之前,所有像素值都归一化为−1到1的范围。

数据集被随机划分为训练集和测试集。训练集是在FTransGAN[17]使用的数据集的基础上扩展而来的,并添加了从开放访问网站收集来的额外字体文件,最终的训练集包含818种字体,每种字体包含大约993个中文字符。然后从训练集中的每种字体中选择52个字符作为样式参考图像,并从52个字符中随机选择6个图像作为风格图像输入到风格编码器中。而内容图像选择微软雅黑作为标准字体,标准字体仅用于索引字符类别。测试集由两部分组成,用于测试未知样式和未知内容的泛化能力。一种是将剩余的27种字体作为风格未知内容已知的测试集,另一种则是使用29个未知字符作为内容未知样式已知的测试集。

字体数据集和分区规则的几个示例如图5所示。其中S1是用于网络的训练集;S2是样式已知但内容未知的测试集,其内容没有在训练中出现过;S3是内容已知但风格未知的测试集,它的字体风格没有训练中出现过。

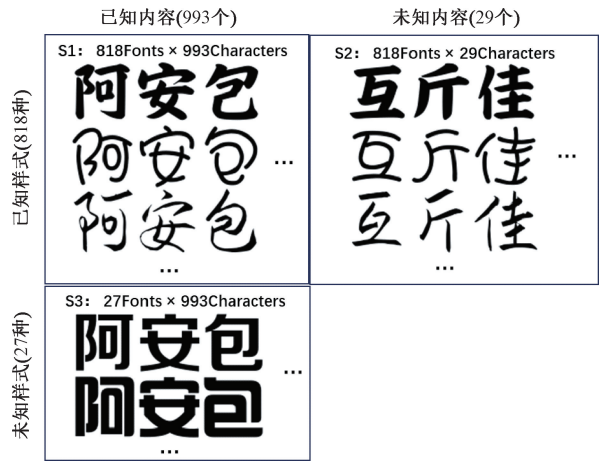

图5 数据集的划分规则

2.2 评估指标

在字体生成任务中没有标准的评估指标来衡量生成的图像。因此,本文从像素级和感知级两个角度对模型进行综合评价。

对于像素级评估,本文主要使用MAE、SSIM和MS-SSIM来计算生成图像与真实图像之间相同位置的距离。

MAE:平均绝对误差是衡量生成图像与真实图像之间差异的一种指标。当平均绝对误差的值越小,生成的图像与目标字体越接近。其公式如下所示:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-x_i| \tag{17}$$

式中:y 是真实图像,x 是生成的图像。

SSIM:结构相似性指标是从亮度相似性、对比度相似性和结构相似性3个方面来评估生成图像的质量。当SSIM的值越大,生成的图像与目标字体越相似。其公式如式(18)所示。

$$SSIM(x,y)=l(x,y)c(x,y)s(x,y) \tag{18}$$

其中,$l(x,y),c(x,y),s(x,y)$ 分别表示两幅图像的平均亮度、平均对比度和相关系数之间的距离。

MS-SSIM:多尺度相似性指数是一种将图像分解为多个尺度,每个尺度代表不同级别的细节。然后在每个尺度上计算SSIM指数,并将它们组合起来提供更全面的结构相似性度量。

其公式如式(19)所示。

$$MS-SSIM(x,y)=[l_M(x,y)]^{\alpha_M}\cdot\prod_{j=1}^{M}[c_j(x,y)]^{\beta_j}[s_j(x,y)]^{\gamma_j} \tag{19}$$

其中,$l_M(x,y)$ 表示仅在尺度 M 上计算两幅图像的平均亮度,$c_j(x,y)$ 和 $s_j(x,y)$ 分别表示在第 j 个尺度上计算两幅图像的平均对比度和相关系数间的距离。指数 α_M、β_j 和 γ_j 用于调整不同成分的相对重要性。

像素级的评估指标有时会与人类的直觉产生矛盾。因此,还需要从感知级的层次来评估不同方法的模型。通过结合像素级和感知级两个层次的指标来全面评估所有的模型。对于感知级指标,本文主要使用mFID和准确率来评估,使生成的图像更加符合人类的感官。

mFID:感知级评估提出了一种通过计算真实图像与生成图像在特征空间之间的距离(fréchet inception distance, FID)来评估模型的方法。Liu等[21]在FID的基础上通过计算每个目标类的平均FID,将其修改为条件版本(mFID)。本文在自建的数据集上训练了两个ResNet-50网络,分别对生成的字体图像进行内容和风格分类。然后从内容和风格两个方面计算mFID的值。mFID的值越小,生成的图像越逼真。

Acc:准确率是指生成正确图像占总图像数量的比例。此处主要用ResNet-50网络对生成图像进行内容和风格分类,同时报告了内容和风格的top-1准确率。准确率在数学上的定义如下:

$$Acc=\frac{T}{T+F} \tag{20}$$

其中,T 表示识别正确的数量,F 表示识别错的数量。

2.3 实验设置

由于项目是在pixpix框架上对FTransGAN进行改进的,因此实验中的基本设置都遵循pixpix设置。在实验中,设置 $\lambda_1=100$,$\lambda_c=\lambda_s=1$,风格图像数量 $k=6$。生成器 G、内容判别器 D_c 和风格判别器 D_s 都使用正态初始化。同时使用Adam优化器对模型进行20个epochs的训练,其中 $\beta_1=0.5$,$\beta_2=0.999$,前10个epochs的学习率为0.000 2,后10个epochs的学习率呈线性衰减,批次大小设

置为 64。

2.4　对比方法

本文选择了 5 种有监督的字体生成方法与所提出的方法进行定性和定量比较。5 种模型的相关介绍如下：

EMD：Zhang 等[22]提出了一种广义的风格迁移框架（separating style and content for generalized style transfer，EMD）。它利用内容和风格的条件依赖性，将图像分解为独立的风格和内容表示，最后将内容和风格特征简单的拼接在一起输入到解码器中生成目标图像。

DFS：Zhu 等[23]提出了一种通过加权深度特征（deep feature similarity，DFS）的方法来对目标进行风格化。它把目标字符与参考字符间的内容特征的距离计算为相似矩阵，并进一步整合为单个风格特征的权重，最后将风格与内容特征混合并输入解码器中生成与参考字符相同风格的字符。

LSG-FCST：Li 等[24]提出一种分层相似引导（layer similarity guiding few-shot chinese style transfer，LSG-FCST）的方法将低级语义到高级语义的内容和风格特征输入到注意力机制中，然后整合不同层的相似特征生成风格化特征。该方法在 FTransGAN 框架上改进的，但它没有使用上下文感知模块和层注意力模块。

MF-Net：Zhang 等[25]在 FTransGAN 的基础上提出了一种新颖的语言复杂性感知跳跃连接（multilingual font generation method，MF-Net），并将其应用于未知语言。这种跳跃连接可以自适应地调整内容的结构信息，从而使生成的图像与目标字体具有相似的风格。

FTransGAN[17]：FTransGAN 设计了上下文感知模块和层注意力模块捕捉局部和全局的风格特征，然后将捕获的风格特征输入解码器中生成字体图像，使跨语言生成任务成为可能。

为了公平比较，选择微软雅黑作为内容字体，并在 2.1 节的数据集上训练所有的对比方法。

2.5　对比实验

本文从像素级和感知级两个角度评估提出的方法与其他方法在未知内容和未知风格的性能。表 1 和表 2 分别是不同方法在未知内容数据集和未知风格数据集的实验结果。从像素级到感知级，提出的方法总体上优于其他对比模型。

表 1　未知字符图像的评估结果

Methods	MAE	SSIM	MS-SSIM	Acc(c)	Acc(s)	mFID(c)	mFID(s)
EMD	**0.121 0**	0.481 9	0.449 9	0.785 7	0.177 2	162.13	654.68
DFS	0.165 8	0.348 6	0.280 5	0.747 3	0.048 2	284.38	736.08
MF-Net	0.232 6	0.231 5	0.198 6	0.981 8	0.032 5	167.78	802.01
LSG-FCST	0.127 7	0.479 2	0.467 6	0.965 0	0.494 8	57.84	332.49
FTransGAN	0.122 4	0.497 9	0.491 0	0.977 2	0.600 5	47.40	290.71
Ours	**0.121 0**	**0.502 7**	**0.497 8**	**0.983 1**	**0.618 9**	**44.02**	**280.83**

表 2　未知风格图像的评估结果

Methods	MAE	SSIM	MS-SSIM	Acc(c)	Acc(s)	mFID(c)	mFID(s)
EMD	**0.132 7**	0.452 3	0.454 3	0.834 5	0.122 8	235.76	597.53
DFS	0.183 0	0.297 5	0.283 3	0.829 1	0.037 7	301.60	597.43
MF-Net	0.217 1	0.269 1	0.273 0	0.997 0	0.054 6	165.69	698.51
LSG-FCST	0.143 7	0.458 9	0.476 0	0.999 3	0.384 4	104.48	229.77
FTransGAN	0.138 8	0.474 2	0.489 4	0.993 7	0.465 5	101.96	190.48
Ours	0.137 2	**0.477 1**	**0.494 2**	**0.999 4**	**0.486 8**	**97.48**	**170.27**

在未知字符图像下，EMD 的 MAE 指标与提出的方法持平，但 EMD 在感知级指标中的表现并不理想，这说明 EMD 生成的图像与人类的感知是相互冲突的。MF-Net 的内容准确率略高于本文方法，但在其他指标的表现中尚有欠缺。对于 DFS、LSG-FCST 和 FTransGAN，提出的方法在像素级和感知级指标中都明显优于它们。定量结果表明该方法在未知内容的字体上具有很好的泛化能力。

本文在未知风格图像上的相似性指标 MAE、SSIM 和 MS-SSIM 分别为 0.121 0、0.502 7、0.497 8；在感知级评估中，内容和风格上的准确率分别为 0.983 1 和 0.618 9，内容和风格上的 mFID 得分为 44.02 和 280.83。

在未知风格数据集的设置下，EMD 的相似性指标 MAE 略高于提出的方法，其他的指标则表现出非常差的性能。与其他的对比模型相比，提出的方法在像素级和感知级评估中均表现出最好的性能。结果表明，提出的方法可以在未知风格的字体中提取出独特的风格表示，并生成与目标字体相同的高质量图像。

本文在未知风格数据集上的相似性指标 MAE、SSIM

和 MS-SSIM 分别为 0.137 2、0.477 1、0.494 2;在感知级评估中,内容和风格上的准确率分别为 0.999 4 和 0.486 8,内容和风格上的 mFID 得分为 97.48 和 170.27。

为进一步分析所有的对比方法,本文对不同的方法生成的结果进行可视化处理。图 6 随机展示了每种方法生成的 3 种不同风格的字体,包括艺术字体和手写体。对于一些笔画较浅的字体,EMD 无法生成图像,并且在高难度的字体上表现出模糊、笔画缺失等情况。DFS 的性能是不稳定的,它容易丢失一些笔画或者某些风格容易失真,导致生成的字体视觉质量较差。MF-Net 能捕捉到风格字体的特征信息,但是生成的图像中保留了大量的内容特征。虽然 LSG-FCST 和 FTransGAN 能够生成稳定的字符,但它们不能准确描述样式的细节,部分生成结果缺失样式细节。相比之下,即使在结构复杂的字符上,本文的方法仍然能从风格图像中学习到重要的风格特征,并将风格特征转移到目标字体中。

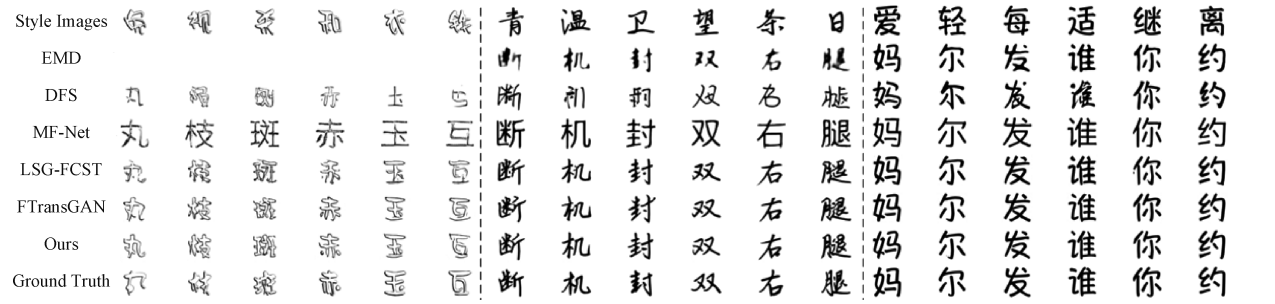

图 6 不同方法的可视化结果

2.6 消融实验

为验证多层次通道注意力网络的有效性,本文将其分为两部分来实现,即风格感知注意力模块和基于风格感知的多头注意力模块。表 3 和 4 的后 4 行展示了各模块的定量结果。其中,Cat 表示对内容和风格编码器的输出不进行任何处理,只是在通道上简单地拼接并输入到解码器中生成结果。SM 表示仅使用风格感知注意力模块。MH 表示将风格感知注意力模块设计成多层次的注意力形式,即多层次的注意力机制。Full model 表示完整模型的定量结果,即在多层次的注意力机制的基础上借鉴 CBAM 的思想所实现的结果,其完整的框架结构如图 1 所示。

表 3 未知字符图像的消融结果

Methods	MAE	SSIM	MS-SSIM	Acc(c)	Acc(s)	mFID(c)	mFID(s)
w/o L_c & L_s	**0.117 2**	0.494 0	0.466 2	0.791 6	0.244 7	153.31	616.05
w/o L_1	0.183 0	0.331 8	0.260 1	0.754 2	0.274 2	179.03	448.28
$\lambda_1=1$	0.166 2	0.371 8	0.320 6	0.861 2	0.332 8	117.81	412.11
$\lambda_1=10$	0.140 4	0.442 6	0.421 9	0.959 3	0.444 1	66.41	353.01
$\lambda_1=50$	0.125 2	0.489 0	0.480 5	0.982 4	0.566 7	47.13	301.59
Cat	0.127 8	0.479 4	0.468 3	0.973 8	0.516 1	50.60	323.27
SM	0.122 6	0.497 0	0.488 5	0.976 4	0.614 0	47.35	284.95
MH	0.122 0	0.499 6	0.494 4	0.980 7	0.598 2	45.67	292.29
Full model	0.121 0	**0.502 7**	**0.497 8**	**0.983 1**	**0.618 9**	**44.02**	**280.83**

表 4 未知风格图像的消融结果

Methods	MAE	SSIM	MS-SSIM	Acc(c)	Acc(s)	mFID(c)	mFID(s)
w/o L_c & L_s	**0.130 7**	0.460 3	0.462 4	0.883 4	0.209 0	209.09	558.25
w/o L_1	0.200 6	0.276 9	0.243 7	0.790 3	0.303 7	304.44	266.57
$\lambda_1=1$	0.181 9	0.334 1	0.321 2	0.933 9	0.368 6	189.18	226.17
$\lambda_1=10$	0.157 5	0.409 0	0.422 5	0.994 0	0.464 4	125.54	184.78
$\lambda_1=50$	0.141 5	0.461 6	0.477 4	0.996 8	**0.509 8**	130.44	173.66
Cat	0.145 1	0.447 7	0.458 0	0.998 2	0.459 6	105.04	184.50
SM	0.143 4	0.456 8	0.467 7	0.997 7	0.430 8	101.98	199.71
MH	0.137 7	0.476 8	0.493 8	0.997 9	0.490 5	98.94	177.82
Full model	0.137 2	**0.477 1**	**0.494 2**	**0.999 4**	0.486 8	**97.48**	**170.27**

评估结果见表 3 和表 4。当加入了风格感知注意力模块后，模型的性能得到了大幅度的提升，说明风格感知注意力模块可以捕获重要的区域风格特征。把风格感知注意力模块设计成多层次的注意力形式后，模型的性能得到进一步的提升，说明不同接受域的风格感知模块可以学习关注不同关系的风格特征，并提升了生成图像的质量。通过结合不同层次的风格特征来实现全局特征的获取，这种结合局部和全局特征的方式能更好地从图像中学习细微的风格表示。其中，完整的模型表现出最好的性能。

为了直观的进行比较，对消融实验进行了可视化。不同消融模型的生成图像如图 7 所示。由于没有进行任何处理就把内容和风格编码器的输出简单拼接输入到解码器中，生成的字符出现了一些伪影，例如第二行的“尾”、“丹”、“抹”。当加入了风格感知注意力模块后，减轻了伪影出现的情况，但仍然会出现少许的笔画断连。在将风格感知注意力模块设计成多层次的注意力形式后，模型的性能得到进一步提升，生成的图像质量得到提高，但个别字符存在笔画不清晰的情况。完整的模型展示了最好的生成结果，生成的字符没有出现以上情况，平均池化有助于以类似于 CBAM 的方式生成通道注意力映射。

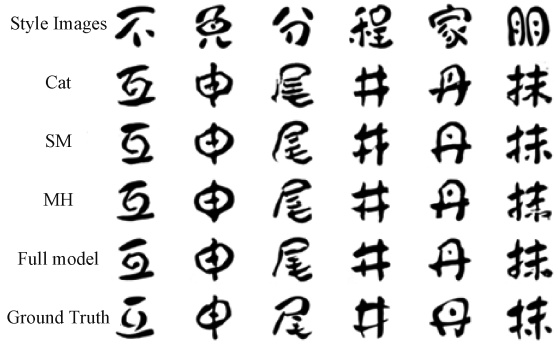

图 7　消融实验的可视化结果

结合表 3、4 和图 7 可知，风格感知注意力模块和多层次注意力形式对最终结果都有积极的影响，说明多层次的通道注意力网络对字体生成任务是有效的，它很大程度上改善了生成图像的质量。

2.7　参数设置分析

在训练过程中，从给定的内容图像 I_c 和风格图像 I_s 中生成对应样式的目标字形 y_c。损失包括 3 部分：L_1 损失、风格损失 L_s 和内容损失 L_c，完整的目标函数如式(16)所示。在式(16)中，将目标函数的权重 λ_1，λ_c，λ_s 分别设置为 100，1，1。为验证权重 λ_1，λ_c，λ_s 设置的合理性，本文在 L_1 损失和对抗损失（L_c 和 L_s）之间设置了不同的参数。

表 3 和 4 报告了不同超参数在内容和风格数据集上的定量结果。当目标函数中没有对抗损失时，模型在像素级指标中表现出较好的性能，感知级指标则表现出比较差的性能。当训练中没有 L_1 损失时，模型表现出最差的性能。当超参数 λ_1 逐渐增大时，模型的性能也随着变好。

不同超参数的模型的可视化结果如图 8 所示。从图中可发现，为目标函数分配不同的权重对模型的性能都有不同程度的影响。当目标函数中没有对抗损失时，生成的字符出现模糊的情况，不能生成完整的字符。当没有 L_1 损失时，生成的字符出现多余的笔画。当 L_1 的权重逐渐增大时，图像的质量也随着改善。其中完整的模型（$\lambda_1=100$，$\lambda_c=\lambda_s=1$）表现出较好的视觉质量，符合人类的感知。

图 8　不同超参数的模型的可视化结果

综上所述，在训练过程中将目标函数的权重 λ_1，λ_c，λ_s 分别设置为 100，1，1。

3　结　　论

本文提出一种新的少样本字体生成模型，该模型通过风格感知模块对风格图像的空间位置和风格特征建模，并利用建模的对应关系寻找重要的局部风格特征；然后进一步将风格感知模块设计成多层次的注意力形式来获取不同接受域的风格特征，并把不同接受域的风格特征聚合成全局特征。此外，本文还利用平均池化和多层感知机为每个特征通道分配一个重要程度，使网络能获得重要的风格特征。最后，在包含 845 种字体的自建数据集上进行了大量的对比实验和消融实验。实验结果表明，提出的模型在像素级和感知级评估方面都明显优于其他方法。

提出的方法可以减少专业字体设计的时间，减轻个人购买版权字体的负担，为公司和个人节省了大量的时间和金钱。其次，它还可以应用于其他的场景，例如图像翻译。此外，该方法还能应用于未知语言字体生成和多种语言之间的转换。

然而，该方法仍有许多挑战需要解决，例如目前的模型只能输出大小为 64×64 的图像。因此，需要将模型的输出改成以矢量形式存储的字体文件。此外，模型需要一定的配对数据集来监督训练，在后续的工作中还要将模型改成无监督训练。

参考文献

[1]　陈慧娴，吴一全，张耀. 基于深度学习的三维点云分析

方法研究进展[J]. 仪器仪表学报,2023,44(11):130-158.
[2] 王莉,董鹏豪,王瞧,等. 基于改进 ResNet18 的干香菇等级识别[J]. 国外电子测量技术,2024,43(1):117-125.
[3] 陈法法,董海飞,邓斌,等. 改进 U_Net 网络的钢结构表面锈蚀图像分割方法[J/OL]. 电子测量与仪器学报:1-9[2024-04-13]. http://kns.cnki.net/kcms/detail/11.2488.TN.20240301.1016.020.html.
[4] 钱燕芳,王敏. 基于新的风格损失函数的图像风格转换方法[J]. 电子测量技术,2019,42(4):70-73.
[5] 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, 2017: 2223-2232.
[6] CHANG B, ZHANG Q, PAN S, et al. Generating handwritten chinese characters using cyclegan[C]. 2018 IEEE winter Conference on Applications of Computer Vision(WACV), IEEE, 2018: 199-207.
[7] HUANG G, LIU Z, VAN D M L, et al. Densely connected convolutional networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.
[8] XIE Y, CHEN X, SUN L, et al. Dg-font: Deformable generative networks for unsupervised font generation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 5130-5140.
[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,2017: 1501-1510.
[10] 刘宇,丁阳,FATIMAHBINTI K,等. 联合内容和风格表示的无监督字体生成网络[J/OL]. 计算机辅助设计与图形学学报:1-13[2024-04-13]. http://kns.cnki.net/kcms/detail/11.2925.TP.20240314.1519.022.html.
[11] ZENG S, PAN Z. An unsupervised font style transfer model based on generative adversarial networks[J]. Multimedia Tools and Applications, 2022, 81(4): 5305-5324.
[12] 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, 2019: 5880-5888.
[13] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1125-1134.
[14] CHEN J, JI Y, CHEN H, et al. Learning one-to-many stylised Chinese character transformation and generation by generative adversarial networks[J]. IET Image Processing, 2019, 13(14): 2680-2686.
[15] CHOI Y, CHOI M, KIM M, et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8789-8797.
[16] 王存睿,丁阳,刘宇,等. 融合笔画语义和注意力机制的汉字字体生成算法[J]. 计算机辅助设计与图形学学报,2022,34(8):1229-1237.
[17] LI C, TANIGUCHI Y, LU M, et al. Cross-language font style transfer[J]. Applied Intelligence, 2023: 1-15.
[18] YU D, FU J, MEI T, et al. Multi-level attention networks for visual question answering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4709-4717.
[19] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision(ECCV), 2018: 3-19.
[20] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral Normalization for Generative Adversarial Networks[J]. ArXiv Preprint, 2018, ArXiv: 1802.05957.
[21] 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,2019: 10551-10560.
[22] ZHANG Y, ZHANG Y, CAI W. Separating style and content for generalized style transfer[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8447-8455.
[23] ZHU A, LU X, BAI X, et al. Few-shot text style transfer via deep feature similarity[J]. IEEE Transactions on Image Processing, 2020, 29: 6932-6946.
[24] LI Y, LIN G, HE M, et al. Layer similarity guiding few-shot Chinese style transfer[J]. The Visual Computer, 2023: 1-14.
[25] ZHANG Y, MAN J, SUN P. MF-Net: a novel few-shot stylized multilingual font generation method[C]. Proceedings of the 30th ACM International Conference on Multimedia, 2022: 2088-2096.

作者简介

邱燕波,硕士研究生,主要研究方向为计算机视觉、图像处理、字体生成等。

储开斌,硕士,教授,硕士生导师,主要研究方向为智能制造技术、机器人等。

E-mail:ckb910@163.com

张继,硕士,副教授,硕士生导师,主要研究方向为计算机视觉、图像处理等。

冯成涛,博士,讲师,硕士生导师,主要研究方向为惯性视觉里程计等。