E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP3-05　　**文献标识码**:A　　**文章编号**:1006-8961(2025)07-2378-11

论文引用格式:Hou D H and Zhu L Q. 2025. IconFormer: an icon generation model based on CNNs and Transformer. Journal of Image and Graphics, 30(7):2378-2388(候冬辉, 竺乐庆. 2025. 基于CNNs和Transformer的图标生成模型IconFormer. 中国图象图形学报, 30(7):2378-2388)[DOI: 10.11834/jig.240570]

基于CNNs和Transformer的图标生成模型IconFormer

候冬辉,竺乐庆*

1. 浙江工商大学计算机科学与技术学院,杭州　310018; 2. 浙江省大数据与未来电子商务技术重点实验室,杭州　310018

摘　要:**目的**　图标自动生成可以提高软件图形用户界面设计的效率,现有的图标自动生成方法存在多样性不足、生成过程复杂以及输入要求较高等问题,限制了生成结果的自由度和创新性。本文提出一种基于Transformer的高效且灵活的图标生成方法,该方法只需提供任意一对内容图标和风格图标,即可生成一幅新的具有特定风格的图标图像。**方法**　提出一个图标生成模型IconFormer,网络结构中包括一个VGG(Visual Geometry Group)特征编码器、一个基于卷积神经网络(convolutional neural network,CNN)的风格编码器、一个Transformer多层解码器和一个CNN解码器,并用内容损失、风格损失、一致性损失和梯度损失组成的综合损失来优化网络模型。**结果**　为了评估所提出的图标生成模型,构建了包含43 741个图标样本的数据集,在该数据集上对IconFormer模型进行训练和评估,并在相同条件下与先进的相关方法进行对比和分析。评估结果表明,本文的IconFormer生成的图标在颜色和结构上更为完整,而其他相关方法则一定程度出现了内容缺失、风格化不足和背景着色的情况,IconFormer在内容差异和梯度分数等量化指标上也明显优于其他模型。消融实验进一步表明了本文所构建的IconFormer模型各个创新点对图标生成过程所起的正向作用。**结论**　所提出的图标生成模型IconFormer,结合了卷积神经网络和Transformer模型的优点,可以快速高效地生成具有不同风格的高质量图标。

关键词:图标生成;图像风格迁移;卷积神经网络(CNN);Transformer;自注意力机制

IconFormer: an icon generation model based on CNNs and Transformer

Hou Donghui, Zhu Leqing*

1. School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou　310018, China;

2. Zhejiang Key Laboratory of Big Data and Future E-Commerce Technology, Hangzhou　310018, China

Abstract: **Objective** Icons are essential components for the graphical user interface (GUI) design of software or web sites because they can quickly and directly convey their meaning to users through visual information, improving the usability of software and websites. However, manually creating a large number of icon images with a consistent style and a harmonious color scheme is a labor-exhaustive and time-expensive procedure. Moreover, professional artists are required to do this job. Therefore, researchers have explored methods for automatically generating icons by using deep learning models to improve the efficiency of GUI design in software. Several state-of-the-art icon generation methods have been proposed in recent years. However, some of these methods based on generative adversarial networks suffer from the problem of insuffi-

收稿日期:2024-10-11;**修回日期**:2024-12-20;**预印本日期**:2024-12-27

*通信作者:竺乐庆　zhuleqing@zjgsu.edu.cn

基金项目:浙江省"尖兵""领雁"科技计划项目(2023C01042)

Supported by:Zhejiang Province Leading Geese Plan(2023C01042)

cient diversity in the generated icons, while some of these methods require users to provide initial icon sketches or color prompts as auxiliary inputs, increasing the complexity of the generation process. Therefore, this study proposes a novel icon generation method based on Transformer and convolutional neural network (CNN), with which new icons can be generated based on given pair of content icon and style icon. In this manner, icons can be generated more efficiently and flexibly than in previous methods with better quality. The proposed model in this study, called IconFormer, can effectively establish the relationship between content and style through Transformer and avoid the problems of missing local detail information of the content and insufficient stylization. **Method** This study proposes an icon generation model, called IconFormer, based on deep neural networks. The network architecture is composed of a feature encoder based on VGG, a style encoder based on CNNs, a multilayer Transformer decoder, and a CNN decoder. The style encoder is designed to discover more style information from style features. The Transformer decoder achieves a high degree of integration between content encoding and style encoding. To train and test the proposed icon generation model, this study collects a high-quality dataset that contains 43 741 icon images, comprising icons of different styles, categories, and structures. The icon dataset is organized into pairs, with each pair containing a content icon and a style icon. The dataset is divided into a training set and a testing set, following a ratio of 9∶1. The content and style features are first extracted from the input content icon and style icon with the ImageNet pretrained VGG19 encoder, and then the style features are further encoded into style key $\boldsymbol{K}$ and style value $\boldsymbol{V}$ with the style encoder. Subsequently, the content features as $\boldsymbol{Q}$, style key $\boldsymbol{K}$, and style value $\boldsymbol{V}$ are inputted into the multilayer Transformer decoder for feature fusion. Finally, the fused features are decoded into a stylized new icon with the CNN decoder. A new loss function integrated by content loss, style loss, identity loss, and gradient loss is adopted to optimize network parameters. **Result** The proposed IconFormer is evaluated on the icon dataset and compared with previous state-of-the-art methods under the same configuration. These state-of-the-art methods include AdaIN(adaptive instance normalization), ArtFlow, StyleFormer, StyTr^2(style transfer transformer), CAP-VSTNet(content affinity preserved versatile style transfer network), and S2WAT(strips window attention Transformer). The experimental results suggest that the icons generated by the proposed IconFormer are more complete in color and structure than those generated by the previous methods. The icons generated by AdaIN, ArtFlow, and StyleFormer demonstrate content loss and insufficient stylization in different extents. StyTr^2 cannot effectively distinguish the primary structure from background information of an icon, and most of the background of its generated icons are colorized. The quantitative analysis results show that the proposed IconFormer outperforms previous methods in terms of content and gradient differences. AdaIN results in the highest content difference, indicating that this method exhibits content loss, while ArtFlow presents the highest style difference, indicating that this method cannot effectively stylize content icons. Several ablation experiments are conducted to verify the effectiveness of the feature encoder, style encoder, and loss function definition in the icon generation process. The result shows that the VGG feature extractor, style encoder, and integrated loss function with gradient loss have positive effects on the resulting icons. Additional experiments are conducted to generate a set of icons with a unified style, and the results show that IconFormer is extremely convenient to generate a set of icons with a consistent style, harmonious colors, and high quality. **Conclusion** The icon generation model, IconFormer, based on CNNs and Transformer proposed in this study combines the advantages of CNNs and Transformers, and thus can generate new icons with high quality and efficiency, saving time and labor cost for the GUI design of software or websites.

Key words: icon generation; image style transfer; convolutional neural network (CNN); Transformer; self-attention mechanism

0　引　言

图标是软件或网页的图形用户界面(graphical user interface,GUI)不可或缺的一个重要组成部分。在用户界面设计中,一个简洁、精美的图标不仅可以吸引用户的注意,还能够帮助用户快速识别功能,提高用户的使用效率。人工制作网站网页和应用程序中风格一致、配色和谐的一套主题图标需要专业的美工人员耗费较长的时间来完成,代价高昂。由深度生成模型自动化生成大规模、风格一致的图标数据集可以提高软件行业的生产效率、节约软件开发

成本，对软件产业有重要意义。

已经有研究人员在图标自动生成或图标风格迁移等方向做了一些探索和研究。Yang等人（2021）和Chen等人（2022）对原本用于生成人脸图像的StyleGAN模型进行优化，以适应图标生成任务，从而生成一系列丰富的图标。Sun等人（2019）和Han等人（2020）在条件生成对抗网络中探究利用多个不同的辨别器约束生成图标的颜色、结构等信息。与上述基于对抗生成网络（generative adversarial network，GAN）的方法不同，Li等人（2022）提出一种图标草图着色的方法，将风格和结构特征分离，使用归一化流网络生成多样化和逼真的彩色图标。

另外一种尝试是采用图像风格迁移的方法，在保留原始图标内容的基础上，将风格转换为与另一张图标的风格相一致，从而生成一张新颖的图标。基于优化的图像风格迁移方法（Gatys等，2016）通过不断迭代实现风格化图像的结构接近内容图像，而风格分布接近风格图像，获得特定风格和内容主题的图标，但是该方法效率较低。Justin等人（2016）采用感知损失函数训练前馈神经网络，明显提高了风格迁移的速度，但是只能实现单一风格的迁移。Stylebank（Chen等，2017）将多种风格合并到一个模型中，但它仍然无法处理未经训练的风格。Huang和Belongie（2017）在传统的编码器—解码器框架中加入了自适应实例归一化（adaptive instance normalization，AdaIN），在特征空间中调整内容图像特征的均值和方差，使其与风格图像的均值和方差相匹配，实现了实时的、任意风格的图像风格迁移。廖远鸿等人（2023）通过分层提取与合成风格向量，实现生成的图像具有不同风格迁移强度，并生成细节更丰富的结果图像。AdaAttN（adaptive attention normalization）（Liu等，2021）在卷积神经网络（convolutional neural network，CNN）模型的基础上，加入了新的注意力和归一化模块，使得模型更加关注浅层特征和局部特征统计信息，实现了内容图像和风格图像的特征有效融合。风格迁移模型CAP-VSTNet（content affinity preserved versatile style transfer network）（Wen等，2023）由一个可逆残差网络和一个无偏线性变换模块组成，不仅可以保留内容细节，还避免引入冗余信息，从而实现更好的风格化效果。

上述基于卷积神经网络的方法在处理输入图像之间的全局信息和长距离依赖时表现出一定的局限性。Transformer架构中的自注意力机制能够有效捕捉全局上下文信息，减少风格迁移过程中内容的泄露并实现了无偏的风格化。StyleFormer（Wu等，2021）利用Transformer捕捉长距离依赖，同时实现风格化图像与内容图像语义内容的一致性和精细的风格多样性。StyTr^2（style transfer transformer）（Deng等，2022）则由两个Transformer编码器和一个Transformer解码器组成，采用Transformer可以捕捉精细内容特征的优势，有效地消除了内容泄露和风格偏差的问题。孙梅婷等人（2023）提出一个结合CNN和Transformer的混合网络，同时引入一个判别网络，提高了生成的风格化图像真实感。S2WAT（strips window attention Transformer）（Zhang等，2024）是一种层次化的Transformer架构，通过计算不同窗口形状，从而改善其中的注意力机制。Puff-Net（Zheng等，2024）设计了两个独立的提取器用于提取内容和风格特征，然后通过只有编码器的Transformer模型进行风格融合，显著降低了计算成本，并实现了较好的风格迁移效果。Transformer也得以有效应用于图像生成（Carlier等，2020；Chen等，2021）、文本—图像生成（Reddy等，2021；Wu等，2023）和基于文本的图标生成任务中（Wu等，2023；Lin等，2024）。

综上所述，目前针对图标自动生成任务存在如下挑战：1）缺少一个内容和风格丰富多样的高质量图标数据集用于训练基于深度学习的图标自动生成模型；2）图标通常结构简单、形状相对规则，有些具有立体效果或渐变颜色，使得图标与真实图像在视觉表现和处理要求上存在显著差异，现有的图像生成模型用于生成图标效果较差；3）生成模型需要有效地建立内容图标和风格图标之间的联系，在改变图标风格的同时，还必须保留其内容信息，在确保图标的语义一致性方面仍存在一定的困难。为了解决上述问题，本文首先收集并建立了一个具有不同风格、内容和场景的高质量图标数据集，并提出一种在传统的编码器—解码器框架中引入基于CNN的风格编码器和基于Transfomer的多层解码器的图标风格迁移模型IconFormer，其中风格编码器用于提取风格编码，Transformer多层解码器用于图标内容和风格编码融合。采用图标风格迁移的方法，参考一张风格图标的风格，改变内容图标的风格，从而设计出一套风格一致的图标集，实现个性化主题的定制。

本文的贡献概括如下：1）创建了一个涵盖不同风格、内容和结构的高质量图标数据集，该数据集共包括来自48个场景的43 741个图标；2）提出一个采用风格迁移的图标生成方法，提取内容图标的结构特征与风格图标的视觉风格，将其相融合，生成全新的图标；3）提出一个命名为IconFormer的图标生成模型，整合了CNN的局部特征提取能力和Transformer的全局特征捕捉能力，生成高质量丰富多彩的新图标。

1　IconFormer图标生成模型

1.1　IconFormer整体网络结构

本文所提出的IconFormer图标生成模型整体网络结构如图1所示，整个模型由特征编码器E、风格编码器（style encoder，SE）、Transformer解码器（Transformer decoder，TD）、图标解码器D构成。首先，内容图标$\boldsymbol{I}_c$和风格图标$\boldsymbol{I}_s$分别输入到编码器E中，编码器通过多层卷积神经网络提取图像特征，输出内容特征$\boldsymbol{Z}_c$和风格特征$\boldsymbol{Z}_s$，在本文中编码器E采用ImageNet预训练的VGG19（Visual Geometry Group 19）；其次，风格特征$\boldsymbol{Z}_s$由风格编码器SE提取风格键矩阵$\boldsymbol{K}$和风格值矩阵$\boldsymbol{V}$，内容特征$\boldsymbol{Z}_c$作为查询矩阵$\boldsymbol{Q}$。接着，$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$作为序列输入到Transformer解码器中，对内容编码和风格编码进行融合，输出风格化的特征序列$\boldsymbol{Z}_{cs}$；最后，风格化的特征$\boldsymbol{Z}_{cs}$由图标解码器D解码出风格化后的新图标。

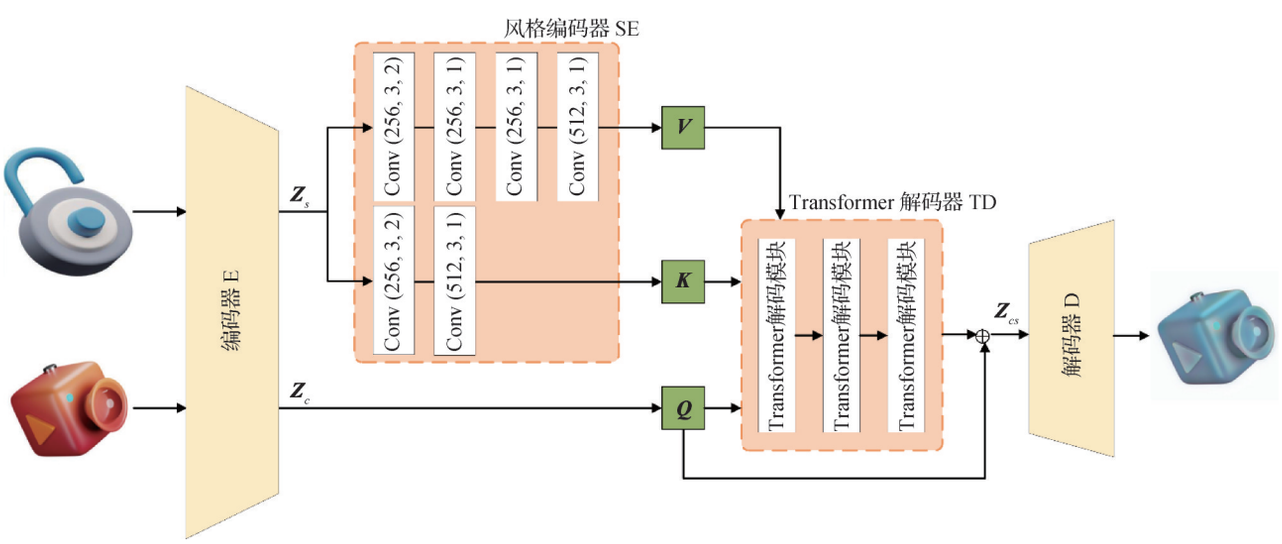


图1　IconFormer模型网络结构

Fig. 1　Network architecture of the IconFormer

1.2　风格编码器SE

风格编码器SE旨在从风格特征$\boldsymbol{Z}_s$中获取更多丰富的风格信息，并生成一组包括风格键$\boldsymbol{K}$和风格值$\boldsymbol{V}$的风格序列编码。具体而言，风格特征$\boldsymbol{Z}_s$输入到风格编码器中后，经过两个并行分支进行处理：第1个分支用于提取风格值编码，其结构上包括一个步长为2的卷积层和3个步长为1的卷积层；第2个分支用于提取风格键编码，由一个步长为2和一个步长为1的卷积层组成。其中，步长为2的卷积层用于减小特征图的空间尺寸，获取更具全局性的风格信息，而步长为1的卷积层则在保持特征图空间尺寸的基础上，进一步细化和增强风格特征的表示能力。

如图1所示，风格编码器的每个卷积层的形式为Conv(C_{out}, k, s)，其中C_{out}是输出通道数量，k是卷积核的大小，s是步长。风格编码器能够有效提取风格特征，这一过程提高了风格信息的表达能力，为下一步的内容和风格融合阶段提供丰富和细致的风格信息。

1.3　Transformer解码器TD

Transformer解码器TD在保留原始内容结构信息的同时，能够实现内容和风格的高度融合和转换。

原始风格特征$\boldsymbol{Z}_s$由风格编码器SE再次提取后，生成了一组包括风格键矩阵$\boldsymbol{K}$和风格值矩阵$\boldsymbol{V}$的风格序列编码。与此同时，原始内容特征$\boldsymbol{Z}_c$用做内容查询矩阵$\boldsymbol{Q}$。Transformer解码器从风格键$\boldsymbol{K}$和风格值$\boldsymbol{V}$中进一步提取风格的高级特征，并将其与内容查询矩阵$\boldsymbol{Q}$进行融合，生成风格化的内容特征序列。具体而言，首先将查询矩阵$\boldsymbol{Q}$、键矩阵$\boldsymbol{K}$和值矩阵$\boldsymbol{V}$作为序列输入到Transformer解码器中。如图1所示，Transformer解码器是一个3层网络结构，其中每一层结构如图2，包括两个多头自注意力模块和一个前馈神经网络。

多头自注意力通过多个并行的注意力头捕捉不同的风格特征和内容特征，这些不同的特征通过通道连接后再由线性变换生成最终多头注意力的输出，具体为

$$F_{\mathrm{MSA}}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=Concat\left(\boldsymbol{H}_1,\cdots,\boldsymbol{H}_n\right)\boldsymbol{W}_o \tag{1}$$

式中，$\boldsymbol{H}_i=\mathrm{softmax}\left[\dfrac{(\boldsymbol{Q}\boldsymbol{W}_i^Q)(\boldsymbol{K}\boldsymbol{W}_i^K)^{\mathrm{T}}}{\sqrt{d_k}}\right](\boldsymbol{V}\boldsymbol{W}_i^V)$，$\boldsymbol{W}_i^Q$，$\boldsymbol{W}_i^K$，$\boldsymbol{W}_i^V$为需要学习的权重矩阵，$d_k$是缩放因子，用于防止内积过大导致梯度消失。$n$是注意力头的数量，$\boldsymbol{W}_o$是参数矩阵。输入的$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$根据两次多头注意力运算，可以更好地融合风格特征和内容特征，具体为

$$\boldsymbol{Y}_1=F_{\mathrm{MSA}}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)+\boldsymbol{Q} \tag{2}$$

$$\boldsymbol{Y}_2=F_{\mathrm{MSA}}\left(\boldsymbol{Y}_1,\boldsymbol{K},\boldsymbol{V}\right)+\boldsymbol{Y}_1 \tag{3}$$

经过两次多头注意力运算后输出的序列$\boldsymbol{Y}_2$以残差连接的方式输入到前馈神经网络，输出风格化序列$\boldsymbol{Y}_{cs}$，具体为

$$\boldsymbol{Y}_{cs}=FFN\left(\boldsymbol{Y}_2\right)+\boldsymbol{Y}_2 \tag{4}$$

式中，$FFN(\boldsymbol{Y})=\max(0,\boldsymbol{Y}\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2$是一个包含两层全连接层的前馈神经网络，在每一层应用了层归一化(Carion等，2020)。

为了将Transformer输出的高维特征映射到图像空间，本文采用了一个3层结构的卷积神经网络图标解码器D对风格化序列进一步解码和细化，每一层包括3×3的卷积、ReLU(rectified linear unit)激活函数和2倍上采样操作。最终，解码器生成尺寸与内容图像一致且通道数量为3的高质量风格化的新图标图像。

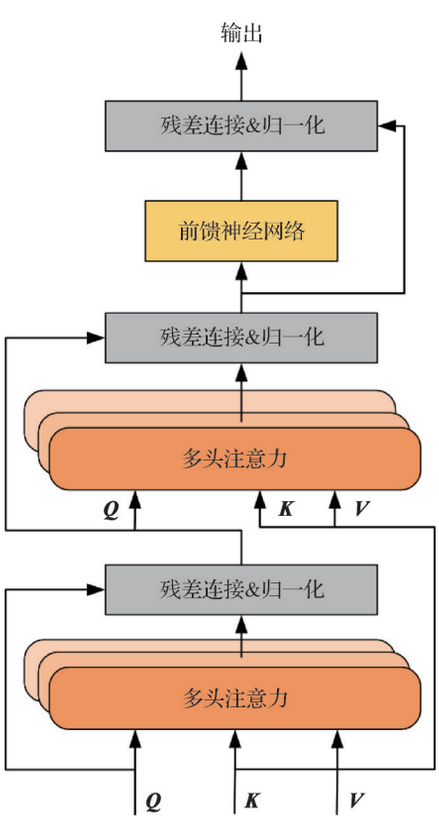


图2　Transformer解码器网络结构

Fig. 2　Network structure of Transformer decoder

1.4　损失函数设计

本文所提出的IconFormer的损失函数由广泛使用在风格迁移任务中的内容损失、风格损失以及一致性损失(Park和Lee，2019)和梯度损失组成，通过优化总损失函数不断调整模型参数。

内容损失确保风格化图标在内容结构上与原始内容图标接近一致，采用内容损失L_c衡量内容图标$\boldsymbol{I}_c$和生成图标$\boldsymbol{I}_{cs}$之间的内容差异，具体为

$$L_c=\frac{1}{N_l}\sum_{i=0}^{N_l}\left\|\phi_i(\boldsymbol{I}_{cs})-\phi_i(\boldsymbol{I}_c)\right\|_2 \tag{5}$$

风格损失L_s确保生成图标在风格分布上与参考风格图标相似，计算风格图标和生成图标的均值和方差，衡量风格图标$\boldsymbol{I}_s$和生成图标$\boldsymbol{I}_{cs}$之间的风格分布差异，具体为

$$L_s=\frac{1}{N_l}\sum_{i=0}^{N_l}\left\|\mu(\phi_i(\boldsymbol{I}_{cs}))-\mu(\phi_i(\boldsymbol{I}_s))\right\|_2+\left\|\sigma(\phi_i(\boldsymbol{I}_{cs}))-\sigma(\phi_i(\boldsymbol{I}_s))\right\|_2 \tag{6}$$

式中，ϕ_i表示预训练的VGG19提取的第i层的特征，$\mu(\cdot)$和$\sigma(\cdot)$分别表示提取特征的均值和方差，N_l是层数。

一致性损失用于帮助图标生成模型学习更丰富、更准确的内容和风格特征。通过将内容图标(或

风格图标)同时作为内容和风格输入到模型中提取内容和风格特征来重建结果图标$\boldsymbol{I}_{cc}$(或$\boldsymbol{I}_{ss}$),结果图标应与原始图标$\boldsymbol{I}_c$(或$\boldsymbol{I}_s$)尽可能保持一致,因此,计算了两个一致性损失项来衡量$\boldsymbol{I}_c$和$\boldsymbol{I}_{cc}$及$\boldsymbol{I}_s$和$\boldsymbol{I}_{ss}$之间的差异,具体为

$$L_{id1} = \left\| \boldsymbol{I}_{cc} - \boldsymbol{I}_c \right\|_2 + \left\| \boldsymbol{I}_{ss} - \boldsymbol{I}_s \right\|_2 \tag{7}$$

$$L_{id2} = \frac{1}{N_l} \sum_{i=0}^{N_l} \left\| \phi_i(\boldsymbol{I}_{cc}) - \phi_i(\boldsymbol{I}_c) \right\|_2 + \left\| \phi_i(\boldsymbol{I}_{ss}) - \phi_i(\boldsymbol{I}_s) \right\|_2 \tag{8}$$

梯度损失计算内容图标和生成图标之间的梯度差异,以保持生成图标的各区域颜色一致性,通过应用Sobel卷积核计算水平和垂直方向的梯度得到,具体为

$$L_{gd} = \frac{1}{N} \sum_{i=1}^{N} ((Conv(\boldsymbol{I}_c, \boldsymbol{K}_x) - Conv(\boldsymbol{I}_{cs}, \boldsymbol{K}_x))^2 + (Conv(\boldsymbol{I}_c, \boldsymbol{K}_y) - Conv(\boldsymbol{I}_{cs}, \boldsymbol{K}_y))^2 \tag{9}$$

式中,$Conv(\boldsymbol{I}, \boldsymbol{K})$表示图像$\boldsymbol{I}$与卷积核$\boldsymbol{K}$的卷积操作,$\boldsymbol{K}_x$和$\boldsymbol{K}_y$分别是Sobel水平方向和垂直方向的卷积核,$N$是图像中的像素数。

总损失函数L为

$$L = \lambda_c L_c + \lambda_s L_s + \lambda_{id1} L_{id1} + \lambda_{id2} L_{id2} + \lambda_{gd} L_{gd} \tag{10}$$

式中,λ_c、λ_s、λ_{id1}、λ_{id2}和λ_{gd}分别代表各损失的权重。实验表明,当内容损失的权重高于风格损失的权重时,风格化效果会显著减弱;当风格损失的权重高于内容损失的权重时,生成图标会出现内容信息丢失或扭曲,或是产生背景着色现象;此外,实验还表明,L_{id1}的值明显小于L_{id2},过大的梯度损失会对图标的生成结果产生负面影响,因此本文分别设定λ_c、λ_s、λ_{id1}、λ_{id2}和λ_{gd}为经验值1、1、50、1和0.01。

2　实验结果与分析

2.1　实验数据集及设置

本文创建的图标数据集大部分从IconScout、Free3Dicon和六图网搜集而来,包括来自48个类目的43 741个图标,其中包括实验室、图表、人物和支付等种类。与真实感图像组成的数据集相比,图标数据集中的图标是经过人工设计的,涵盖了不同种类、结构和颜色分布,有些具有立体效果、复杂的结构或渐变颜色,数据集中的图标样本丰富多样。

为了训练和评估模型,将创建的图标数据集按9∶1的比例划分为训练集和测试集,其中训练集中有39 367幅图标,测试集中有4 374幅图标。接着,分别将训练集和测试集中的图标随机划分为两部分,其中50%作为内容图标,另外50%作为风格图标。在训练阶段,首先将所有图标图像的尺寸统一缩放为512 × 512像素,再随机裁剪出256 × 256像素的子区域形成训练样本。模型的训练参数设置如下:学习率为0.000 5,所有参数采用Adam优化器(Kingma和Ba,2015)进行优化,优化器的参数β_1和β_2设置为0.5和0.999,批次大小设置为8,在配置有NVIDIA RTX 1080Ti显卡的服务器上进行了160 000次迭代训练。

2.2　测试结果及对比实验

本文提出的IconFormer与AdaIN(Huang和Belongie,2017)、ArtFlow(An等,2021)、StyleFormer(Wu等,2021)、StyTr^2(Deng等,2022)、CAP-VSTNet(Wen等,2023)和S2WAT(Zhang等,2024)6种先进的图像风格迁移方法在所创建的数据集上进行了训练与对比测试。上述6种对比方法均使用其默认参数设置,与本文方法在相同的训练集上进行训练,在相同的测试集上进行了对比测试。

2.2.1　可视化结果与对比

本文所提出的图标生成模型IconFormer可视化结果以及与其他6种相关方法的对比结果如图3所示。由图3(i)可知,本文提出的IconFormer模型能够有效区分出图标的主体与背景,从而生成结构完整、清晰准确的图标图像。模型确保了每个区域的颜色趋于一致,颜色分布展现出丰富的多样性,生成的图标具有高度的可识别性、多样性和视觉吸引力,更能满足实际应用的需求。相比较而言,AdaIN(图3(c))和ArtFlow(图3(d))所生成的图标内容信息严重缺失和扭曲,有明显的颜色渗出现象,方法未能有效保持图标的内容信息。StyleFormer(图3(e))生成的图标则出现风格化程度不足,存在各区域颜色一致性差、有明显噪声的问题,且部分图标的细节部位出现了结构扭曲的情况。StyTr^2(图3(f))未能成功区分图标的主体与背景,导致在图标生成过程中出现了背景着色现象,进而明显降低了生成图标的视觉质量。CAP-VSTNet(图3(g))所生成的图标中,存在明显的内容缺失以及部分图标背景着色的问题。S2WAT(图3(h))所生成的图标整体清晰度不足,同样有图标背景着色的情况出现。

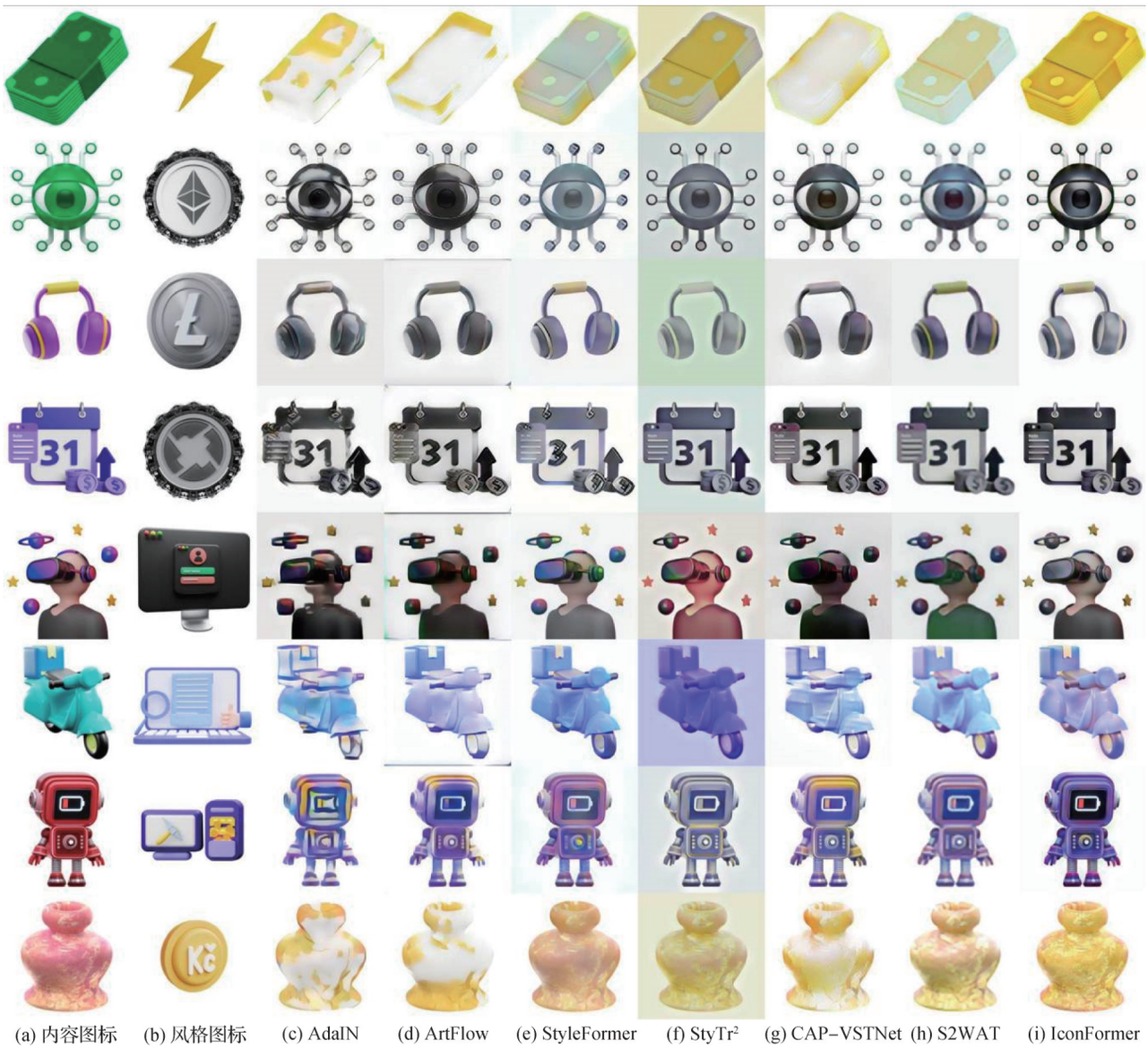


图3　可视化对比结果

Fig. 3　Visualization comparison results ((a) content icons; (b) style icons; (c) AdaIN; (d) ArtFlow; (e) StyleFormer; (f) StyTr^2; (g) CAP-VSTNet; (h) S2WAT; (i) IconFormer)

2.2.2　量化分析与对比

通过计算生成图标与输入内容图标之间的内容差异和梯度差异，与输入风格图标之间的风格差异，作为图标生成质量的3个间接评估指标。内容差异和风格差异越小，表明输入内容和风格保持越好。梯度差异越小，生成图标与输入内容图标相同区域的颜色梯度越接近。对于每一种风格迁移方法，通过将测试集中的2 187幅内容图标与2 187幅风格图标随机组合作为输入，生成2 187幅风格化图标图像。根据式(5)计算内容差异，根据式(6)计算风格差异，并根据式(9)计算梯度差异。量化分析的结果以及与其他6种相关方法的对比如表1所示。由表1可知，本文所提出的IconFormer方法生成的图标在内容差异和梯度差异上都明显低于其他6种相关方法，其风格差异也低于S2WAT、CAP-VSTNet、StyTr^2、StyleFormer、ArtFlow这5种方法，只比AdaIN高。AdaIN方法所生成图标的风格差异最低，但内容差异反而是最高的，这表明该方法在风格迁移过程中不能很好地保护内容结构信息，正如图3可视化结果比较中显示的，该方法出现了明显的内容信息缺失。ArtFlow方法所生成图标的风格差异是最高的，因此该方法在图3中表现出较差的风格化效果。

表 1　量化分析结果及与相关方法的对比

Table 1　Quantitative analysis results and a comparison with related methods

方法	内容差异	风格差异	梯度差异
IconFormer	**0.47**	0.78	**0.15**
S2WAT	0.51	2.53	0.50
CAP-VSTNet	0.52	1.34	0.18
StyTr^2	0.56	2.64	0.17
StyleFormer	0.67	1.11	0.21
ArtFlow	0.82	5.28	0.84
AdaIN	2.58	**0.70**	0.44

注：加粗字体表示各列最优结果。

2.2.3　风格化时间性能比较

表 2 展示了不同的风格迁移方法在 256 × 256 像素和 512 × 512 像素两种图像分辨率下风格化单幅图标所需要的时间。所有方法在相同的实验环境下，使用配置有一张 NVIDIA RTX 1080Ti 显卡的服务器对测试集中的 2 187 幅内容图标和风格图标进行随机结合，测试风格化所需要的平均时间。

表 2　不同方法风格化图标的时间性能比较

Table 2　A comparison with related methods in time performance of stylizing an icon

/s

方法	256 × 256 像素	512 × 512 像素
IconFormer	0.047	0.186
S2WAT	0.042	0.171
CAP-VSTNet	0.063	0.192
StyTr^2	0.058	0.277
StyleFormer	0.015	0.029
ArtFlow	0.071	0.236
AdaIN	**0.014**	**0.021**

注：加粗字体表示各列最优结果。

表 2 的结果显示，AdaIN 模型在两种分辨率下均表现出最短的处理时间，这可以归因于 AdaIN 模型简单，仅由一个编码器和解码器组成。尽管 AdaIN 模型在时间效率上具有优势，但其相对简单的网络结构难以处理较为复杂的图标。本文所提出的 IconFormer 在运行时间上并非最优，但能在生成图标的质量和时间性能之间取得较好的折中。

2.3　消融实验

2.3.1　基于VGG19的编码器作用评估

为了验证 IconFormer 中以 VGG19 为基础的编码器对内容图标和风格图标对提取特征的影响，本文对比了两种特征提取方法的表现：一种是采用 VGG19 和风格编码器 SE，另一种是采用 Transformer 编码器。可视化结果如图 4 所示，可以看出采用 Transformer 编码器的模型所生成的图标出现了颜色缺失现象，在部分区域出现了杂色（图 4(c)）。相对而言，采用 VGG19 和风格编码器的模型表现出对颜色和细节的控制能力，很少出现颜色丢失和不必要的杂色情况，生成的图标质量明显优于前者（图 4(d)）。

(a) 内容图标　(b) 风格图标　(c) Transformer　(d) VGG19

图 4　不同编码器可视化结果对比

Fig. 4　Visualized results comparison for different encoders ((a) content icons; (b) style icons; (c) Transformer; (d) VGG19)

量化分析结果如表 3 所示，结果显示了采用 VGG19 和风格编码器的模型在风格差异略优于采用 Transformer 编码器的模型，但在内容差异方面则显著高于后者。因此，仅使用 Transformer 模型中的解码器，结合 VGG19 和风格编码器的模型更适合图标生成任务。

2.3.2　风格编码器作用评估

为了进一步研究风格编码器 SE 在模型中的有效性，本文对所提出的 IconFormer 模型进行了一系列的消融测试。首先将风格编码器从模型中移除，

表3　不同编码器的量化分析结果

Table 3　Quantitative analysis results of different encoders

编码器	内容差异	风格差异
VGG19和风格编码器	**0.47**	**0.78**
Transformer	0.94	0.80

注：加粗字体表示各列最优结果。

不去提取风格特征Z_s的风格键和风格值编码，而是直接将编码器E提取的风格特征Z_s同时作为K矩阵和V矩阵，内容编码Z_s作为Q矩阵，再将Q、K和V作为序列编码输入到Transformer解码器TD中融合内容编码和风格编码。最终，由图标解码器D生成风格化的图标。消融实验的可视化对比结果如图5所示，由图5可知，该消融模型所生成的图标在部分区域的边缘出现了一些杂色（图5(c)第1、2排），有一些图标的背景出现了明显的着色（图5(c)第3排）。而没有去除风格编码器生成的图标则没有明显出现上述问题（图5(b)），生成图标的视觉质量更高。

(a) 内容图标　(b) IconFormer　(c) 去除SE　(d) 风格图标

图5　风格编码器作用的可视化结果

Fig. 5　Visualized results for the effect of style encoder ((a) content icons; (b) IconFormer with SE; (c) with SE removed; (d) style icons)

表4展示了从IconFormer模型中去除风格编码器的量化分析结果，结果显示IconFormer模型在内容差异和风格差异均优于消融模型。因此，在去除风格编码器SE的情况下，生成的图标在风格一致性上显著下降，证明了风格编码器对模型的性能起到了增强的作用。

表4　风格编码器的量化分析结果

Table 4　Quantitative analysis results of style encoder

编码器	内容差异	风格差异
IconFormer	**0.47**	**0.78**
去除风格编码器	0.74	1.96

注：加粗字体表示各列最优结果。

2.3.3　梯度损失权重评估测试

本文通过调整梯度损失的权重，探讨其对生成风格化图标质量的影响。梯度损失用于计算内容图标与风格化图标之间的梯度差异，目的是在风格迁移的过程中保持内容图标与风格化图标的颜色分布一致性。表5对比了不同梯度损失权重下生成图标与内容图标及风格图标的内容差异和风格差异，其中权重为0时表示不使用梯度损失的基准模型。实验结果显示，加入梯度损失后的生成效果优于基准模型，且在权重为0.01时，模型表现最佳。因此，本文采用梯度损失权重为0.01。

表5　梯度损失的量化分析结果

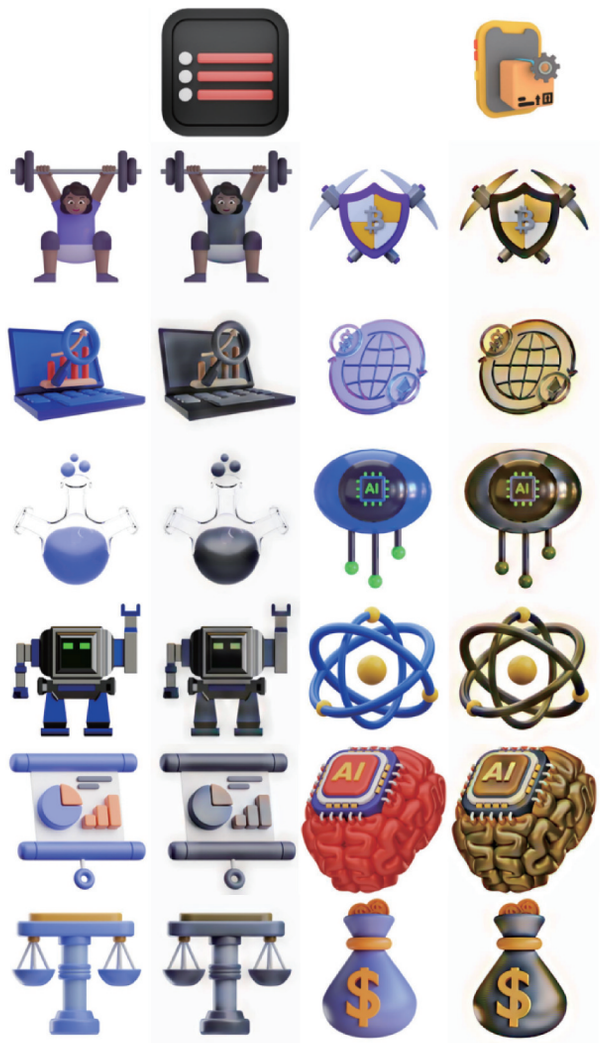
Table 5　Quantitative analysis results for gradient loss

梯度损失	内容差异	风格差异
0	0.65	1.23
5	0.44	1.25
1	0.54	1.19
0.1	0.47	0.86
0.01	**0.43**	**0.78**
0.001	0.53	1.17
0.000 1	0.68	1.28

注：加粗字体表示各列最优结果。

2.4　图标集生成

在实际应用过程中，设计师更倾向于在同一页面中应用一套风格统一的图标集，以确保视觉上的和谐性。如图6所示，在保持一组基准图标风格不变的前提下（即第1行图标），通过选择多个不同内容的内容图标（图6(a)及图6(c)），本文所提出的IconFormer模型能够快速生成一系列风格一致、色彩和谐且高质量的图标集（图6(b)及图6(d)）。这展示了在保持设计风格统一性的同时，实现图标内容的多样化和高效生成，有助于提升用户界面的和谐性和美观度。



(a) 内容图标1　(b) 生成图标1　(c) 内容图标2　(d) 生成图标2

图6　生成风格一致的图标集示例

Fig. 6　Examples of generated icon sets with consistent styles ((a)content icons 1; (b)generated icons 1; (c)content icons 2; (d)generated icons 2)

3　结　论

本文提出一个新颖的图标生成模型IconFormer，为了训练和测试该模型，创建了一个涵盖多种风格、内容和结构的高质量图标数据集，数据集中包含43 741个来自48个类目的图标样本，为图标生成研究提供了一定的数据支撑。本文所提出的基于图像风格迁移的图标生成模型IconFormer，能够根据任意内容图标和风格图标生成具有特定风格和内容的新图标。IconFormer模型结合了卷积神经网络(CNN)和Transformer的优势，能够有效建立内容图标和风格图标之间的联系，显著提升了生成图标的质量。本文所提出的IconFormer在创建的数据集上进行了实验，并与6种相关的图标风格迁移方法进行了对比，可视化结果和量化分析表明，IconFormer所生成的图标具有较高质量，优于其他6种方法所生成的图标。本文通过消融实验进一步分析了VGG编码器、风格编码器SE以及梯度损失对于网络整体性能的贡献。本文所提出的图标生成模型IconFormer能用于生成结构和颜色分布相对复杂的立体效果图标。目前本文只能针对结构固定的图标迁移不同的配色方案，未来的研究方向是对图标的类型进行迁移，生成相同主题但不同类型风格的图标，如将平面图标转换成立体图标。

参考文献(References)

An J, Huang S Y, Song Y B, Dou D J, Liu W and Luo J B. 2021. Artflow: unbiased image style transfer via reversible neural flows//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 862-871 [DOI: 10.1109/CVPR46437.2021.00092]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Carlier A, Danelljan M, Alahi A and Timofte R. 2020. DeepSVG: a hierarchical generative network for vector graphics animation//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #1372

Chen D D, Yuan L, Liao J, Yu N H and Hua G. 2017. StyleBank: an explicit representation for neural image style transfer//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2770-2779 [DOI: 10.1109/CVPR.2017.296]

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. 2021. 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 Y P, Pan Z Y, Shi M, Lu H, Cao Z G and Zhong W C. 2022. Design what you desire: icon generation from orthogonal application and theme labels//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 2536-2546 [DOI: 10.1145/3503161.3548109]

Deng Y Y, Tang F, Dong W M, Ma C Y, Pan X J, Wang L and Xu C S.

2022. StyTr^2: 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]

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]

Han Q R, Zhu W Z and Zhu Q. 2020. Icon colorization based on triple conditional generative adversarial networks//Proceedings of 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP). Macau, China: IEEE: 391-394 [DOI: 10.1109/VCIP49819.2020.9301890]

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]

Justin J, Alexandre 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]

Kingma D P and Ba J. 2015. Adam: a method for stochastic optimization//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR: #6980 [DOI: 10.48550/arXiv.1412.6980]

Li Y K, Lien Y H and Wang Y S. 2022. Style-structure disentangled features and normalizing flows for diverse icon colorization//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11234-11243 [DOI: 10.1109/CVPR52688.2022.01096]

Liao Y H, Qian W H and Cao J D. 2023. MStarGAN: a face style transfer network with changeable style intensity. Journal of Image and Graphics, 28(12): 3784-3796 (廖远鸿, 钱文华, 曹进德. 2023. 风格强度可变的人脸风格迁移网络. 中国图象图形学报, 28(12): 3784-3796) [DOI: 10.11834/jig.221149]

Lin J W, Jiang Z Y, Guo J Q, Sun S Z, Liu T, Yang Z J, Lou J G and Zhang D M. 2024. IconDM: text-guided icon set expansion using diffusion models//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia; : ACM: 156-165 [DOI: 10.1145/3664647.3681057]

Liu S H, Lin T W, He D L, Li F, Wang M L, Li X, Sun Z X, Li Q and Ding E R. 2021. AdaAttN: revisit attention mechanism in arbitrary neural style transfer//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6629-6638 [DOI: 10.1109/ICCV48922.2021.00658]

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]

Reddy M D M, Basha M S M, Hari M M C and Penchalaiah M N. 2021. Dall-e: creating images from text. UGC Care Group I Journal, 8(14): 71-75

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]

Sun T H, Lai C H, Wong S K and Wang Y S. 2019. Adversarial colorization of icons based on contour and color conditions//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 683-691 [DOI: 10.1145/3343031.3351041]

Wen L F, Gao C Y and Zou C Q. 2023. CAP-VSTNet: content affinity preserved versatile style transfer//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 18300-18309 [DOI: 10.1109/CVPR52729.2023.01755]

Wu R H, Su W C, Ma K D and Liao J. 2023. IconShop: text-guided vector icon synthesis with autoregressive transformers. ACM Transactions on Graphics (TOG), 42(6): #230 [DOI: 10.1145/3618364]

Wu X L, Hu Z H, Sheng L and Xu D. 2021. StyleFormer: real-time arbitrary style transfer via parametric style composition//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14598-14607 [DOI: 10.1109/ICCV48922.2021.01435]

Yang H Y, Xue C Q, Yang X Y and Yang H. 2021. Icon generation based on generative adversarial networks. Applied Sciences, 11(17): #7890 [DOI: 10.3390/app11177890]

Zhang C Y, Xu X G, Wang L, Dai Z Y and Yang J. 2024. S2WAT: image style transfer via hierarchical vision transformer using strips window attention//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 7024-7032 [DOI: 10.1609/aaai.v38i7.28529]

Zheng S, Gao P and Zhou P. 2024. Puff-Net: efficient style transfer with pure content and style feature fusion network//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8059-8068 [DOI: 10.1109/CVPR52733.2024.00770]

作者简介

候冬辉, 男, 硕士研究生, 主要研究方向为图像生成和风格迁移。E-mail: hdhui103@163.com

竺乐庆, 通信作者, 女, 副教授, 主要研究方向为图像处理和人工智能技术。E-mail: zhuleqing@zjgsu.edu.cn