基于双路视觉 Transformer 的图像风格迁移

纪宗杏，贝佳*，刘润泽，任桐炜

(南京大学 计算机软件新技术国家重点实验室，南京 210093）

摘　　要：图像风格迁移(Image Style Transfer, IST)旨在根据风格图像调整内容图像的视觉属性，使其保留原始内容的同时呈现出特定风格样式，从而生成具有视觉吸引力的风格化图像。现有代表性方法大多未考虑不同图像域间的编码差异，专注提取图像局部特征而忽视了全局上下文信息的重要性。为此，本文提出了一种新型的基于双路视觉 Transformer 的图像风格迁移方法 ***Bi-Trans***，对内容图像域和风格图像域进行独立编码，提取风格参数向量以离散化表征图像风格，通过交叉注意力机制与条件实例归一化将内容图像标定至目标域风格，从而生成风格化图像。实验结果表明，本文方法无论是内容保留度还是风格还原度均优于现有方法。

关　键　词：图像风格迁移；视觉 Transformer；任意风格化；条件实例归一化；注意力机制

图像风格迁移(Image Style Transfer, IST)旨在将图像渲染为特定风格，同时保留其原始内容语义。虽然，生成式 AI(AI Generated Content, AIGC)也能够利用内容图像与若干个文本提示词自动化生成具有特定视觉语义的精美图像，满足个性化服务与智能性创作需求。然而，不同于 IST 任务定义，即仅调整内容图像的视觉属性而非生成全新图像，图像风格迁移相关的 AIGC 研究[1-3]通常基于隐式扩散模型(Latent Diffusion Model)[4]，存在大规模修改内容图像的问题。此外，由于文本提示词难以组织且文本语义的细微变化将生成截然不同的图像内容，AIGC 无法实现稳定可控的图像生成。相比之下，IST 以内容图像为生成基准并以风格图像为样式参考，其生成图像的可预期性更强，同时也更加符合人们美化既有图像的基本需求。因此，IST 任务仍然具有较大的发展意义与研究价值。

随着深度学习技术在计算机视觉领域的渗透应用，大量基于卷积神经网络(Convolutional Neural Network, CNN)的图像风格迁移方法被提出并取得可观成效，可大致划分为基于优化的方法和基于前馈的方法[5]。

基于优化的方法通过不断迭代更新噪声图像，持续融合内容特征与风格特征，以实现风格化。Gatys等人[6]采用Gram矩阵表征图像风格，基于预训练VGG19网络优化噪声图像，使其呈现内容图像的高层语义，同时具有风格图像的特征相关性，推动了基于优化的图像风格迁移技术研究。然而，低效的迭代过程使此类方法无法满足实际应用需求。

基于前馈的方法通过训练图像转换网络，从而在推理阶段经一次前馈即可生成风格化图像。Johnson等人[7]引入感知损失，并以无监督方式训练了一个转换网络，实现了实时图像风格迁移。Ulyanov等人[8]提出了一个多尺度纹理生成网络，以完全前馈方式迁移艺术风格，并进一步验证了实例归一化对于提升风格化质量的有效性。Lin等人[9]设计了一个两阶段图像风格化过程，即先迁移全局样式生成一张风格化草图，再在高分辨率下修补其局部纹理细节，从而能实时生成高质量迁移图像。上述方法虽能高效执行图像风格化，但不具备泛化能力，即需要重新训练模型以适配新风格图像。Ghiasi等人[10]通过训练一个预测网络提取面向特定风格的特征缩放参数，并通过条

件实例归一化方式将内容特征映射至风格图像域，得到风格化图像。HuangX等人[11]以无参数方式调整内容特征的整体分布，使其匹配风格特征的通道均值与标准差。同样地，Li等人[12]对内容特征进行白化与着色，以使其协方差匹配至风格特征。Dae等人[13]通过风格注意力机制将局部风格样式灵活融合至内容特征图中，从而达到风格化效果。Liu等人[14]提出了一个新的注意力和归一化模块，能够在每个点的基础上自适应地执行注意力归一化，同时考虑了深层特征与浅层特征，从而在保持内容结构的同时更好地转移风格样式。Chandran等人[15]通过在风格转移过程中预测卷积核和偏置传递风格图像的统计特征和结构信息，使得风格迁移不仅仅局限于全局特征，而是能够更细致地处理局部结构。

随着视觉Transformer的提出，其全局信息整合能力有效弥补了CNN局部表征的短板，使模型更加关注内容语义的形状与结构。Deng等人[16]首次将视觉Transformer引入到图像风格迁移任务中，使用两个Transformer图像编码器表征内容语义和纹理样式，并通过Transformer图像解码器交叉融合不同图像域特征，进而转换生成风格化图像。Zhang等人[17]引入条形窗口注意力用于整合局部信息，并集成水平与垂直方向上的长距离依赖关系，从而消除风格化图像中的网格伪影。Wang等人[18]引入两种注意力机制，基于自注意力编码图像块序列间的相关性，并通过交叉注意力将风格样式自适应合成至内容图像中。Zhang等人[19]引入了一种新的边缘损失，用于增强内容的细节，在风格特征过度渲染导致的模糊情况下改善图片的清晰度。Feng等人[20]提出了一种新颖的组合式Transformer自编码器，分别处理高相关性和低相关性的特征块，能够有效地处理和保留内容图像的结构和语义信息。然而，现有基于视觉Transformer的IST方法大多存在形状偏向，忽视了对于目标纹理及色彩分布的精准刻画。

为提升风格迁移质量，本文提出了一种基于双路视觉 Transformer 的图像风格迁移方法，利用Transformer架构捕获图像长距离依赖关系以建模全局上下文信息，并引入 Transformer 风格参数提取器对图像风格进行离散化表示，最终在解码架构中将内容嵌入配准对齐至风格特征，从而实现图像风格化。实验结果表明，本文方法优于现有最先进方法，能实现高质量任意图像风格迁移。

本文主要贡献包括以下两个方面：(1)本文提出了一种基于双路视觉 Transformer 的图像风格迁移方法，针对内容图像域和风格图像域进行独立编码，捕获图像长距离依赖关系以建模全局上下文信息，兼具较高的内容保真度与风格还原度；(2)本文引入了一个 Transformer 风格参数提取器，通过交叉注意力机制提取风格参数，将图像风格离散化表示为一组关键特征，提升了风格表征刻画的丰富性和准确性; (3)本文在 Transformer 图像解码阶段显式度量风格分布特征，基于预测得到的风格特征分布参数反归一化网络中间嵌入，从而基于特征分布匹配实现特定风格渲染。

1 卷积神经网络和视觉 Transformer

1.1 卷积神经网络

LeCun 等人[21]在 1998 年提出了用于手写数字识别的卷积神经网络 LeNet-5。Krizhevsky 等人[22]提出的 AlexNet 在 ImageNet 竞赛中取得的出色表现，奠定了卷积神经网络在计算机视觉领域的地位。

得益于 CNN 强大的非线性表征学习能力，图像风格迁移技术得到蓬勃发展。基于 CNN 的图像风格迁移方法通常按编码-解码架构划分网络逻辑，其中编码器大多基于预训练图像分类网络，其表征能力直接决定了风格呈现质量。这类方法需堆叠大量卷积池化层以获取充足的感受野，从而将低层视觉特征重组为高层抽象语义。与此同时，网络量级和复杂度得到同步提升，且特征分辨率不断降低，进而无法精准刻画内容语义及纹理样式[16]。此外，这类方法未考虑不同图像域间的编码差异，将基于自然图像数据集训练所得的编码器直接应用于编码风格图像，从而引入表征误差。如文献[23,24]中所述，CNN 还存在明显的纹理偏向，不利于保留内容图像的原始语义结构。

1.2 视觉 Transformer

Dosovitskiy 等人[25]在 2020 年提出的 ViT(Vison Transformer, ViT)，它将图片裁剪为 16 个切块，并在切块序列的每个切块上直接应用 Transformer，在图像分类任务中取得优越性能的同时，还具有很强的扩展性。Dosovitskiy[25]证明了当拥有足够多的数据时，ViT 的表现会超过 CNN，推动了后续视觉 Transformer 的研究。

随着 Transformer 在计算机视觉领域的成功应用，图像风格迁移任务被重新赋予生命力。与 CNN 不同，视觉 Transformer 依赖注意力机制在网络浅层即可捕获图像全局信息，建模图像块间的长距离依赖关系，与人类通过形状辨认物体的视觉感知特性相似。然而，视觉 Transformer 无法有效表征各图像块内部的像素相关性，即图像局部特征刻画能力较差，存在形状偏向[24]。现有代表性方法大致遵循原始视觉 Transformer 结构设计，即先基于自注意力机制进行图像编码，再通过计算内容编码与风格编码间的交叉注意力实现特征融合，从而将内容图像映射至目标风格域。这类方法无法准确塑造风格图像中的色彩分布与显著样式，当参考风格中存在精细纹理时，风格整体呈现质量可能不及最先进 CNN 方法。

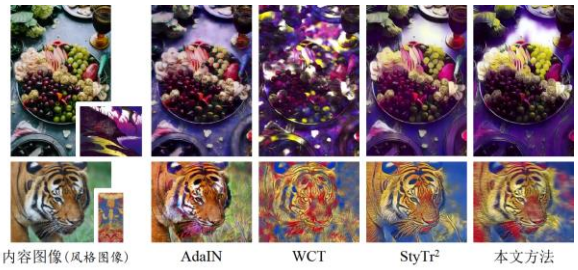

图 1 本文方法与代表性方法的风格化效果对比

Fig.1 Comparison of stylization results between the method in this paper and some representative methods

1.3 小结

作为图像风格迁移任务中的基本构成要素，形状与纹理分别对应内容和风格，任一偏向类型均不利于实现高质量图像渲染。图 1 对比了本文方法与三种现有代表性图像风格迁移方法的风格化效果。其中，AdaIN[11]和 WCT[12]基于 CNN，而 StyTr²[16]基于视觉 Transformer。由此可见，AdaIN[11]由于忽视内容图像域与风格图像域间的编码差异，其风格表征能力较差，参考样式几乎不可感知；WCT[12]存在严重的纹理偏向，注重局部纹理刻画而削弱了风格化图像的整体性；StyTr²[16]具有较强的形状偏向，缺乏纹理感，其色彩分布也与参考风格间存在细微偏差。

2 基于双路视觉 Transformer 的图像风格化

本文提出了一种基于双路视觉Transformer的图像风格迁移方法，称作*Bi-Trans*，主体架构依托于现有基线方法StyTr²[16]。*Bi-Trans*将图像风格迁移任务流程划分为风格参数预测和风格化图像块序列生成务。给定内容图像 $\boldsymbol{I}_{\mathrm{c}} \in \mathbb{R}^{H\times W\times 3}$ 和风格图像 $\boldsymbol{I}_{\mathrm{s}} \in \mathbb{R}^{H\times W\times 3}$，最终生成风格化图像 $\boldsymbol{I}_{\mathrm{cs}} \in \mathbb{R}^{H\times W\times 3}$。式中：$H$、$W$ 分别表示图像高度和图像宽度。Wei 等人[24]表明CNN存在纹理偏向，而视觉Transformer存在形状偏向。为增强图像编码的纹理偏向，*Bi-Trans*没有采用StyTr²[16] 图像划分编码阶段中的线性投影方式，而是将其更换为预训练VGG19网络，分别提取 $\boldsymbol{I}_{\mathrm{c}}$、$\boldsymbol{I}_{\mathrm{s}}$ 在relu4_1层上的特征编码 $\mathcal{F}_{\mathrm{c}}^{4\_1} \in \mathbb{R}^{(H/m)\times(W/m)\times C}$ 和 $\mathcal{F}_{\mathrm{s}}^{4\_1} \in \mathbb{R}^{(H/m)\times(W/m)\times C}$，从而将输入图像划分为 $L=(H/m)\times(W/m)$ 个 $m\times m$ 大小的图像块，得到相应的图像块特征序列 $\boldsymbol{\mathcal{E}}_{\mathrm{c}}=\{\mathcal{E}_{\mathrm{c}}^{1},\mathcal{E}_{\mathrm{c}}^{2},\cdots,\mathcal{E}_{\mathrm{c}}^{L}\}\in\mathbb{R}^{L\times C}$ 和 $\boldsymbol{\mathcal{E}}_{\mathrm{s}}=\{\mathcal{E}_{\mathrm{s}}^{1},\mathcal{E}_{\mathrm{s}}^{2},\cdots,\mathcal{E}_{\mathrm{s}}^{L}\}\in\mathbb{R}^{L\times C}$。式中：$m=8$ 表示图像块尺寸；$C=512$ 表示特征维度。图2展示了本文方法的整体流程。

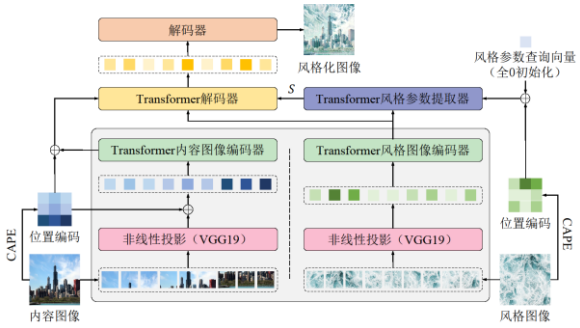

图 2 基于双路视觉 Transformer 的图像风格迁移方法流程图

Fig.2 Flow chart of image style transfer method based on dual vision transformers

2.1 可学习的位置编码

为编码图像块间的空间分布关系，将内容图像块特征序列 $\boldsymbol{\mathcal{E}}_{\mathrm{c}}$ 输入到 Transformer 内容图像编码器之前，本文先对各内容图像块进行位置编码。本文希望在具有相似语义的内容图像块上施加一致的风格化效果，且同一图像区域的位置编码应不随图像尺度变化而改变。如文献[16]中所述，图像块是基于语义上下文和内容相关性进行排列的，而非类似于文本所蕴含的自然逻辑。因此，本文选择采用 Deng 等人[16]提出的基于内容感知的位置编码(Content-Aware Positional Embedding, CAPE)，将计算过程定义为：

$$\hat{\boldsymbol{\mathcal{P}}}_{\mathrm{c}} = Conv_{1\times1}(AvgPool_{n\times n}(\boldsymbol{\mathcal{E}}_{\mathrm{c}})), \tag{1}$$

式中：$Conv_{1\times1}$ 表示点卷积运算；$AvgPool$ 表示平均池化操作；$n=18$ 表示平均池化后的特征图分辨率；$\hat{\boldsymbol{\mathcal{P}}}_{\mathrm{c}} \in \mathbb{R}^{n\times n\times C}$ 表示内容图像位置编码。具体

而言，先通过平均池化将 $\mathcal{E}_c$ 分辨率降低至 $n\times n$；再通过点卷积操作实现跨通道语义信息融合；最后通过双线性差值将 $\hat{\mathcal{P}}_c$ 上采样至与 $\mathcal{E}_c$ 尺寸相同，得到内容图像块位置编码序列 $\mathcal{P}_c=\{\mathcal{P}_c^1,\mathcal{P}_c^2,\cdots,\mathcal{P}_c^L\}\in\mathbb{R}^{L\times C}$。该位置编码方式具有语义一致性和尺度不变性等优点，更适用于图像风格迁移任务。

对于风格图像而言，本文希望在具有相似语义的图像块上提取出一致的风格参数，能有效反映其结构特性，即纹理排列与色调变化。类似地，本文对风格图像进行位置编码，得到 $\hat{\mathcal{P}}_s\in\mathbb{R}^{n\times n\times D}$。式中：$D=768$ 为点卷积运算的输出通道数。

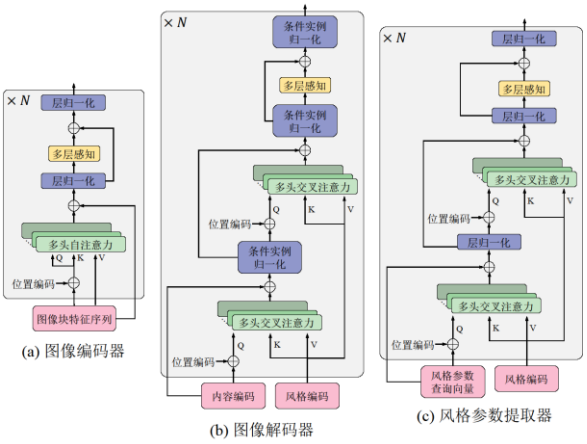

图 3 视觉 Transformer 编解码架构图

Fig.3 The encoder-decoder architecture of Vision Transformer

2.2 Transformer 图像编码器

Wei 等人[24]表明基于 ImageNet 数据集预训练的视觉 Transformer，因存在特定于分类任务和模型结构的形状偏向，不适合用作图像风格迁移任务的特征提取器。此外，如引言中所述，内容图像和风格图像分属不同的图像域，若采用通用编码器将引入特征误差。因此，本文分别为内容图像域和风格图像域重新设计训练了 Transformer 编码器以捕获图像长距离依赖关系，提取领域特定的图像编码。

Transformer 内容图像编码器共堆叠了 N 个编码层，每层包含一个多头自注意力(Multi-head Self-Attention, MSA)模块和一个多层感知(Multi-Layer Perceptron, MLP)模块，其结构如图 3(a)所示。为融合内容图像块的空间位置信息，将 $\mathcal{E}_c$ 与 $\mathcal{P}_c$ 相加，得到内容输入序列 $\boldsymbol{E}_c=\{\mathcal{E}_c^1+\mathcal{P}_c^1,\mathcal{E}_c^2+\mathcal{P}_c^2,\cdots,\mathcal{E}_c^L+\mathcal{P}_c^L\}\in\mathbb{R}^{L\times C}$，并计算其多头自注意力：

$$\begin{aligned}MSA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(&Attn(\boldsymbol{Q}\boldsymbol{W}_1^{\mathrm{q}},\boldsymbol{K}\boldsymbol{W}_1^{\mathrm{k}},\boldsymbol{V}\boldsymbol{W}_1^{\mathrm{v}}),\cdots,\\&Attn(\boldsymbol{Q}\boldsymbol{W}_h^{\mathrm{q}},\boldsymbol{K}\boldsymbol{W}_h^{\mathrm{k}},\boldsymbol{V}\boldsymbol{W}_h^{\mathrm{v}}))\boldsymbol{W}^{\mathrm{o}},\end{aligned}\tag{2}$$

式中：$\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{E}_c$、$\boldsymbol{V}=\mathcal{E}_c$ 分别表示查询向量、键向量和值向量；h 为多头注意力头个数；$\boldsymbol{W}_i^{\mathrm{q}},\boldsymbol{W}_i^{\mathrm{k}},\boldsymbol{W}_i^{\mathrm{v}}\in\mathbb{R}^{C\times(C/h)}$ 为第 i 个注意力头对应的参数矩阵；$\boldsymbol{W}^{\mathrm{o}}\in\mathbb{R}^{C\times C}$ 为输出投影矩阵；$Attn$ 为文献[26]中的注意力操作；$Concat$ 表示维度拼接。本文进一步通过 MLP 模块挖掘内容图像特征的非线性关系，其定义了两层全连接，将注意力输出先升维再降维。本文还对各模块输入输出进行残差连接，经层归一化(Layer Normalization, LN)后再输入至下一模块，最终生成内容编码 $\boldsymbol{\Phi}_c\in\mathbb{R}^{L\times C}$。因此，Transformer 内容图像编码器每层的计算过程可定义为：

$$\begin{aligned}\hat{\boldsymbol{\Phi}}_c&=LN(MSA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{Q}),\\\boldsymbol{\Phi}_c&=LN(MLP(\hat{\boldsymbol{\Phi}}_c)+\hat{\boldsymbol{\Phi}}_c).\end{aligned}\tag{3}$$

类似地，本文构建了 Transformer 风格图像编码器，其结构与内容图像编码器相同，以提取风格输入序列 $\boldsymbol{E}_s=\{\mathcal{E}_s^1,\mathcal{E}_s^2,\cdots,\mathcal{E}_s^L\}\in\mathbb{R}^{L\times C}$ 特征，得到风格编码 $\boldsymbol{\Phi}_s\in\mathbb{R}^{L\times C}$。由于风格化图像中无需保留风格图像的语义结构，$\boldsymbol{E}_s$ 中无需添加风格图像块位置编码。本文进一步对 $\boldsymbol{\Phi}_s$ 及逆行维度变换并计算其通道均值，得到 $\boldsymbol{\Phi}_s^{'}\in\mathbb{R}^{1\times D}$，以作为 Transformer 风格参数提取器输入。

2.3 Transformer 风格参数提取器

受 Ghiasi 等人[10]的启发，为实现任意风格迁移，本文额外设计训练了一个 Transformer 风格参数提取器，其在隐空间中学习各参数分量与风格编码间的相关性，从而使各参数分量侧重反映目标风格的不同特征，实现图像风格拆解并将其离散化表示为若干个关键特征的组合，如颜色特征、形状特征、纹理特征等。

本文将风格参数定义为一个 D 维特征向量 $\boldsymbol{\mathcal{S}}\in\mathbb{R}^{1\times D}$ 并将其初始化为全 0 向量 $\boldsymbol{\mathcal{O}}$。由于风格参数要能反映风格图像的结构信息，本文先对 2.1 节中的风格图像位置编码 $\hat{\mathcal{P}}_s$ 进行通道均值计算，得到 $\mathcal{P}_{\mathcal{S}}\in\mathbb{R}^{1\times D}$；再将其与 $\boldsymbol{\mathcal{O}}$ 相加作为风格参数提取器的查询向量输入，即 $\boldsymbol{Q}=\boldsymbol{\mathcal{O}}+\mathcal{P}_{\mathcal{S}}$。如图 3(c)所示，Transformer 风格参数提取器共堆叠了 N 个提取层，每层包含两个多头交叉注意力(Multi-head Cross-Attention, MCA)模块和一个

MLP 模块。其中，多头交叉注意力模块的键向量和值向量均为风格编码 $\boldsymbol{\Phi}_{\mathrm{s}}$，即 $\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{\Phi}_{\mathrm{s}}$，其计算方式与 MSA 相同。本文将风格参数提取过程定义为：

$$\begin{aligned}\boldsymbol{\mathcal{S}}'' &= LN(MCA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{Q}),\\ \boldsymbol{\mathcal{S}}' &= LN(MCA(\boldsymbol{\mathcal{S}}''+\boldsymbol{\mathcal{P}}_{\mathcal{S}},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{\mathcal{S}}''),\\ \boldsymbol{\mathcal{S}} &= LN(MLP(\boldsymbol{\mathcal{S}}')+\boldsymbol{\mathcal{S}}').\end{aligned} \tag{4}$$

本文进一步对提取出的风格参数 $\boldsymbol{\mathcal{S}}$ 进行降维，以减少冗余特征、增强可理解性，从而将其精简为 $\hat{\boldsymbol{\mathcal{S}}}\in\mathbb{R}^{1\times U}$。式中：$U=100$ 表示降维后风格参数向量维度。

2.4 Transformer 图像解码器

Transformer 图像解码器基于风格编码 $\boldsymbol{\Phi}_{\mathrm{s}}$ 和风格参数 $\hat{\boldsymbol{\mathcal{S}}}$，对内容编码 $\boldsymbol{\Phi}_{\mathrm{c}}$ 进行分解与重组，使其保留语义结构特征并融合风格样式特征。与风格参数提取器相同，Transformer 图像解码器共堆叠了 N 个解码层，每层包含两个 MCA 模块和一个 MLP 模块，最终生成风格化编码 $\boldsymbol{\Phi}_{\mathrm{cs}}\in\mathbb{R}^{L\times C}$，其结构如图 3(b)所示，将计算过程定义为：

$$\begin{aligned}\boldsymbol{\Phi}_{\mathrm{cs}}^{''} &= CIN(MCA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{Q};\hat{\boldsymbol{\mathcal{S}}}),\\ \boldsymbol{\Phi}_{\mathrm{cs}}^{'} &= CIN(MCA(\boldsymbol{\Phi}_{\mathrm{cs}}^{''}+\boldsymbol{\mathcal{P}}_{\mathrm{c}},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{\Phi}_{\mathrm{cs}}^{''};\hat{\boldsymbol{\mathcal{S}}}),\\ \boldsymbol{\Phi}_{\mathrm{cs}} &= CIN(MCA(\boldsymbol{\Phi}_{\mathrm{cs}}^{'})+\boldsymbol{\Phi}_{\mathrm{cs}}^{'};\hat{\boldsymbol{\mathcal{S}}}),\end{aligned} \tag{5}$$

式中：$\boldsymbol{Q}=\boldsymbol{\Phi}_{\mathrm{c}}+\boldsymbol{\mathcal{P}}_{\mathrm{c}},\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{\Phi}_{\mathrm{s}}$。Transformer 图像解码过程即为计算内容图像块与各风格图像块间的相关性，从而根据内容图像块的语义特征选取最适合施加的显著纹理样式。交叉注意力计算有助于建立内容图像域中的语义概念与风格图像域中的纹理样式之间的关联关系，能根据注意力大小决定施加的纹理类型以及纹理施加的强弱程度，从而实现跨图像域风格纹理迁移。

为利用提取到的风格参数 $\hat{\boldsymbol{\mathcal{S}}}$，本文将各模块后的层归一化替换为条件实例归一化(Conditional Instance Normalization，CIN)，通过学习两个缩放平移参数 $\boldsymbol{\gamma}_{\hat{\mathcal{S}}},\boldsymbol{\beta}_{\hat{\mathcal{S}}}\in\mathbb{R}^{1\times C}$ 来调整特征编码的整体分布，从而将内容特征配准对齐至风格特征，将风格化过程定义为：

$$\begin{aligned}CIN(\boldsymbol{x};\hat{\boldsymbol{\mathcal{S}}}) &= \boldsymbol{\gamma}_{\hat{\mathcal{S}}}\left(\frac{\boldsymbol{x}-\mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)+\boldsymbol{\beta}_{\hat{\mathcal{S}}},\\ \boldsymbol{\gamma}_{\hat{\mathcal{S}}} &= \hat{\boldsymbol{\mathcal{S}}}\boldsymbol{W}_{\gamma}+b_{\gamma},\boldsymbol{\beta}_{\hat{\mathcal{S}}}=\hat{\boldsymbol{\mathcal{S}}}\boldsymbol{W}_{\beta}+b_{\beta},\end{aligned} \tag{6}$$

式中：$\mu(\boldsymbol{x})$、$\sigma(\boldsymbol{x})$ 分别是特征编码 x 的通道均值和通道标准差；$\boldsymbol{W}_{\gamma},\boldsymbol{W}_{\beta}\in\mathbb{R}^{100\times C}$ 为投影矩阵；b_{γ}、b_{β} 为偏置项。

本文进一步对 Transformer 解码输出进行维度变换，将其由特征空间重新映射至像素空间。受 Wei 等人[24]和 Deng 等人[16]的启发，为避免引入感知不一致性与边缘伪影，本文并未直接上采样 Transformer 解码输出，而是定义了一个三层 CNN 解码器，对其进行特征提炼并中和其形状偏向，最终生成风格化图像 $\boldsymbol{I}_{\mathrm{cs}}$。其中，CNN 解码器每层通过一个两倍最近邻上采样来扩大编码尺寸，并通过若干个 3×3 卷积来压缩特征通道。

2.5 损失网络

在模型训练阶段，为专注比较风格化图像与内容图像以及与风格图像的感知差异，本文仍采用预训练 VGG19 网络提取图像特征，利用 CNN 的纹理偏向特性促使模型捕获刻画更加精细的风格纹理，从而对图像编码器、风格参数提取器和图像解码器进行参数优化，以使风格化图像尽可能保留原始内容语义，同时呈现出特定纹理样式。本文遵循 Wei 等人[24]的观点，使用感知损失来评估 $\boldsymbol{I}_{\mathrm{cs}}$ 与 $\boldsymbol{I}_{\mathrm{c}}$ 的语义结构相似性，同时衡量其与 $\boldsymbol{I}_{\mathrm{s}}$ 的视觉外观相似性，从而弱化形状偏向，生成更加真实多样的局部纹理细节。

本文定义 relu4_2 为内容损失计算层，提取 $\boldsymbol{I}_{\mathrm{cs}}$、$\boldsymbol{I}_{\mathrm{c}}$ 在其上的特征编码 $\boldsymbol{\mathcal{F}}_{\mathrm{cs}}^{4\_2}$ 和 $\boldsymbol{\mathcal{F}}_{\mathrm{c}}^{4\_2}$，将内容损失定义为图像特征编码间的均方差：

$$\mathscr{L}_{\mathrm{con}}=\frac{1}{|\boldsymbol{\Psi}_{\mathrm{c}}|}\sum_{l\in\boldsymbol{\Psi}_{\mathrm{c}}}\frac{1}{H^{l}W^{l}}\sum_{i=1}^{H^{l}}\sum_{j=1}^{W^{l}}(\mathcal{F}_{\mathrm{cs},ij}^{l}-\mathcal{F}_{\mathrm{c},ij}^{l})^{2}, \tag{7}$$

式中：$\boldsymbol{\Psi}_{\mathrm{c}}=\{4\_2\}$ 为内容层；H^{l}、W^{l} 分别为 l 层上的特征图的高和宽；$\mathcal{F}_{\mathrm{cs},ij}^{l}$ 为特征图 $\boldsymbol{\mathcal{F}}_{\mathrm{cs}}^{l}$ 中 (i,j) 位置的特征值；$|\cdot|$为集合基数。

本文在 relu1_1 到 relu4_1 层上将 $\boldsymbol{I}_{\mathrm{cs}}$、$\boldsymbol{I}_{\mathrm{s}}$ 编码为 $\boldsymbol{\mathcal{F}}_{\mathrm{cs}}$、$\boldsymbol{\mathcal{F}}_{\mathrm{s}}$，并分别计算得到图像编码对应的分布特征 $\{\mu_{\mathrm{cs}},\sigma_{\mathrm{cs}}\}$ 和 $\{\mu_{\mathrm{s}},\sigma_{\mathrm{s}}\}$，将分割损失定义为两对分布参数的均方差之和：

$$\mathscr{L}_{\mathrm{sty}}=\frac{1}{|\boldsymbol{\Psi}_{\mathrm{s}}|}\sum_{l\in\boldsymbol{\Psi}_{\mathrm{s}}}(\mu_{\mathrm{cs}}^{l}-\mu_{\mathrm{s}}^{l})^{2}+(\sigma_{\mathrm{cs}}^{l}-\sigma_{\mathrm{s}}^{l})^{2}, \tag{8}$$

式中：$\boldsymbol{\Psi}_{\mathrm{s}}=\{1\_1,2\_1,3\_1,4\_1\}$ 为风格层；μ_{cs}^{l}、σ_{cs}^{l} 分别表示特征图 $\boldsymbol{\mathcal{F}}_{\mathrm{cs}}^{l}$ 的通道均值和通道标准差。

为提升图像表征性能，本文还额外引入了两个恒等损失项[13]，其能有效限制不必要的图像变动，并要求当内容图像与风格图像相同时，风格化图像应为输入图像本身。本文固定输入为两张内容图像(风格图像)，经双路视觉 Transformer 编解码，最终生成风格化图像 $\boldsymbol{I}_{\mathrm{cc}}(\boldsymbol{I}_{\mathrm{ss}})$。本文计算风格化图像与输入图像间的像素差异以评估其视觉

语义相似性：

$$\mathscr{L}_{\mathrm{id}}^{1}=\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(\boldsymbol{I}_{\mathrm{cc}}^{ij}-\boldsymbol{I}_{\mathrm{c}}^{ij})^{2}+(\boldsymbol{I}_{\mathrm{ss}}^{ij}-\boldsymbol{I}_{\mathrm{s}}^{ij})^{2},\quad(9)$$

式中：$\boldsymbol{I}_{\mathrm{cc}}^{ij}$ 表示图像 $\boldsymbol{I}_{\mathrm{cc}}$ 在 (i,j) 处的像素值。

由于像素损失无法精准衡量图像间的视觉语义相似性，即当存在像素位移时，视觉语义相差较小的图像可能对应较大的像素损失，从而造成模型误判。因此，本文进一步在特征空间中计算风格化图像和输入图像间的编码差异：

$$\mathscr{L}_{\mathrm{id}}^{2}=\frac{1}{\left|\boldsymbol{\Psi}_{\mathrm{id}}\right|}\sum_{l\in\boldsymbol{\Psi}_{\mathrm{id}}}\frac{1}{H^{l}W^{l}}\sum_{i=1}^{H^{l}}\sum_{j=1}^{W^{l}}(\mathcal{F}_{\mathrm{cc},ij}^{l}-\mathcal{F}_{\mathrm{c},ij}^{l})^{2}+(\mathcal{F}_{\mathrm{ss},ij}^{l}-\mathcal{F}_{\mathrm{s},ij}^{l})^{2},\quad(10)$$

式中：$\boldsymbol{\Psi}_{\mathrm{id}}=\{1\_1,2\_1,3\_1,4\_1\}$ 为恒等损失层。

最终，本文将四个分支损失加权求和，得到总训练损失：

$$\mathscr{L}_{\mathrm{tot}}=\lambda_{\mathrm{c}}\mathscr{L}_{\mathrm{con}}+\lambda_{\mathrm{s}}\mathscr{L}_{\mathrm{sty}}+\lambda_{\mathrm{i}}^{1}\mathscr{L}_{\mathrm{id}}^{1}+\lambda_{\mathrm{i}}^{2}\mathscr{L}_{\mathrm{id}}^{2},\quad(11)$$

式中：λ_{c}、λ_{s}、$\lambda_{\mathrm{id}}^{1}$ 和 $\lambda_{\mathrm{id}}^{2}$ 为各分支损失权重（具体取值详见 3.1 节）。

3 实验与结果分析

3.1 数据集与实验设置

本文使用 MSCOCO 数据集[27]和 WikiArt 数据集[28]提供训练所需的内容图像及风格图像。在训练阶段，本文先将输入图像统一缩放至 280×280，再对其随机裁剪，从而控制输入分辨率为 256×256。本文将堆叠层数 N 、注意力头个数 h 分别设定为 3 和 8，并将各分支损失权重 λ_{c}、λ_{s}、$\lambda_{\mathrm{id}}^{1}$ 和 $\lambda_{\mathrm{id}}^{2}$ 设置为 7、10、70 和 1。本文以 4 为批尺寸训练模型共 16 万次迭代，指定 Adam 为优化器并设置其初始学习率为 $5e-4$。针对前 10000 次训练迭代，本文采用学习率预热策略，随后每轮迭代均进行学习率衰减，衰减系数为 $1e-5$。本文实验基于一台配置 Xeon @2.40GHz 处理器、64GB 内存和 Nvidia RTX 3090 24GB 显卡的计算机。

3.2 风格任意性实验

经充分训练，Transformer 风格参数提取器能从任意风格编码中提取出足以表征目标样式的参数向量，并将其应用于 CIN，以标定内容特征。本文从 WikiArt 官网随机下载了八张绘画作品，以测试本文方法对于未知风格的泛化性能。如图 4 所示，可见：

(1) 本文方法能精准刻画任意参考风格，能有效提取并重塑风格图像中的色彩分布以及纹理样式；

(2) 得益于内容感知位置编码，本文方法能在相同语义上产生一致的风格化效果，有效避免了边缘伪影与过度风格化问题；

(3) 本文方法具有较高的内容保真度，风格化图像中原始内容结构与语义信息依旧清晰可辨认。

图 4 本文方法任意风格迁移效果示例

Fig.4 Examples of arbitrary style transfer results of the method in this paper

3.3 对比实验

本文将 *Bi-Trans* 与六个具有代表性的图像风格迁移方法作对比。其中，基于 CNN 的方法包括 Ghiasi[10]、AdaIN[11]和 SANet[13]，而基于视觉 Transformer 的方法包括 StyTr^{2}[16]、S2WAT[17]和 STTR[18]。

3.3.1 定性对比

图 5 展示了本文方法与六个对比方法的风格迁移效果，可见：

(1) StyTr^{2}[16]由于采用简单线性投影划分表征图像块，加之仅在像素空间中编码图像区域相关性，其风格表现力不及本文方法(如行 4)。此外，StyTr^{2}[16]还倾向于将纹理样式集中迁移至显著物体，而使图像背景呈现弱风格化效果(如行 2、行 3)；

(2) S2WAT[17]在不同输入组合下的迁移质量差异较大，由于将注意力计算限制在横向、纵向以及局部窗口内，存在感受野不足问题，整体内容结构及语义边缘因施加纹理样式而被弱化(如行 4、行 5)；

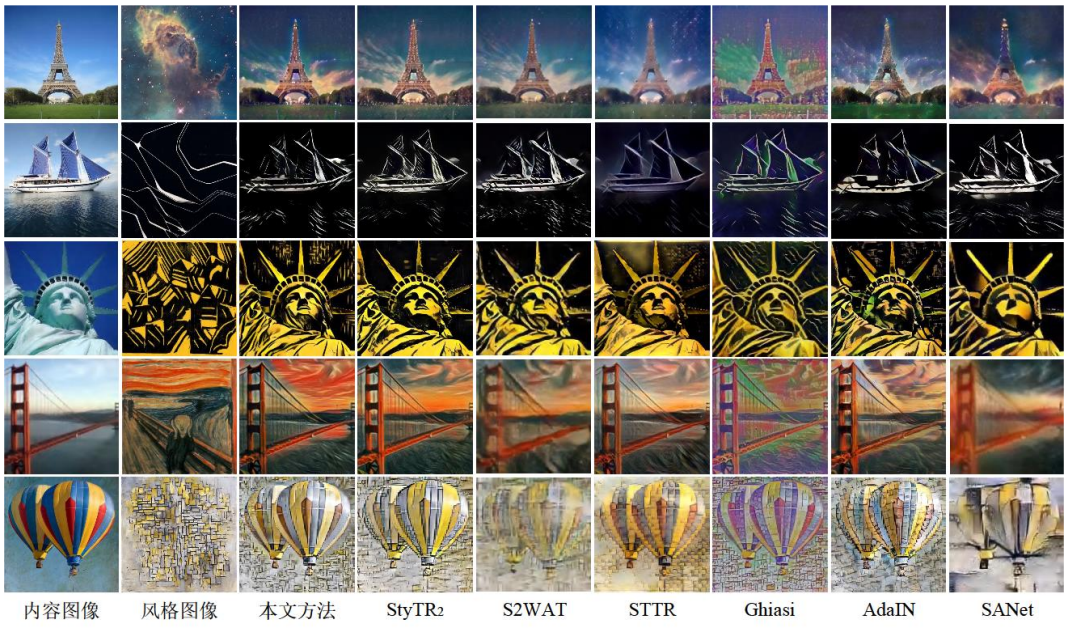

图 5 本文方法与现有方法的风格迁移效果对比

Fig.5 Comparison of style transfer results between method in this paper and existing methods

(3) STTR[18]倾向于保留精细的轮廓结构以尽可能避免内容失真，其风格化效果主要体现在色彩分布，缺乏显著纹理(如行 2、行 4)。此外，STTR[18]对于目标样式的刻画偏离真实纹理(如行 5)，使其风格化质量整体落后于其他基于视觉 Transformer 的方法；

(4) Ghiasi[10]的风格表征能力最差，仅适用于迁移纹理色彩相对单一的简单样式(如行 2、行 3)，对于复杂风格则存在严重的色彩偏差；

(5) 由于 AdaIN[11]直接使用风格特征的分布参数对内容特征进行全局缩放，其存在风格化不完全不充分问题，风格化图像中残留少部分内容伪影(如行 2、行 3)，且对于复杂纹理的重塑能力欠佳(如行 4)；

(6) SANet[13]存在纹理样式施加的合理性问题，目标风格被非均匀地迁移至语义显著物上而产生割裂效果(如行 1)，且同一纹理具有语义异质性(如行 4 中将天空纹理迁移至桥梁)。此外，SANet[13]无法准确迁移复杂风格(如行 5)，特别是当源域和目标域存在较大的视觉结构差异时；

(7) 本文方法与同样基于视觉 Transformer 的方法相比，无论是风格可感知性还是内容可辨识度，均取得可比甚至更优的风格化效果，其能精准刻画显著样式特征，能施加均匀且一致的风格化效果；本文方法与基于 CNN 的方法相比，得益于基础架构的表征能力，更加忠实于原始内容与参考风格，语义纹理间过渡也更加平滑自然。

3.3.2 定量对比

与 Deng 等人[16]相同，本文采用感知损失作为风格化质量评价指标，分别使用式 7 和式 8 来衡量生成图像与输入图像间的内容差异及风格差异，值越小表示迁移质量越高。本文从 MSCOCO 测试集[27]中随机选取了 20 张内容图像，从 WikiArt 官网下载了 20 张不同风格不同流派的艺术画作，从而构成 400 个内容-风格图像对，并计算各方法在所有图像对上的平均内容和平均风格损失。如表 1，2 前三列所示，本文方法在两个评价指标上的平均损失值均最小，进而验证了本文方法相对于现有代表性方法的先进性，表明本文方法能同时兼顾内容语义保留与纹理样式重塑。此外，与上述定性对比结果一致，SANet[13]因执行异质化风格迁移，其平均内容损失最大；

表 1 本文方法与基于CNN的方法平均风格化损失对比

Table 1 Comparison of average stylization loss between method in this paper and different methods based on CNN

方法			
Ghiasi	0.94	2.81	12.51
AdaIN	0.91	1.39	11.63
SANet	0.97	2.88	16.18
本文方法	0.69	1.12	7.31

表 2 本文方法与基于视觉Transformer的方法平均风格化损失对比

Table 2 Comparison of average stylization loss between method in this paper and methods based on vision Transformer

方法	$\mathscr{L}_{\text{con}}\downarrow$	$\mathscr{L}_{\text{sty}}-\mu,\sigma\downarrow$	$\mathscr{L}_{\text{sty}}-$ *Gram*矩阵↓
StyTr^2	0.79	1.35	9.86
S2WAT	0.93	2.64	13.47
STTR	0.79	3.86	27.58
本文方法	0.69	1.12	7.31

STTR[18]因避免内容失真而执行弱风格化，其平均风格损失最大；StyTr^2[16]在两个评价指标上均仅次于本文方法，表明其作为基于视觉 Transformer 的图像风格迁移基线方法的有效性。

除了使用均值、标准差等统计学概念来表征图像风格外，本文还引入风格迁移任务中常用的 Gram 矩阵，其主对角线元素表示各响应层的特征强度，其他元素表示各响应层间的相关系数，从而将图像风格描述为底层特征间的组合关系。同样地，本文在风格层 $\boldsymbol{\Psi}_{\text{s}}$ 上计算基于 Gram 矩阵的风格损失：

$$\mathscr{L}_{\text{sty}}=\frac{1}{\left|\boldsymbol{\Psi}_{\text{s}}\right|}\sum_{l\in\boldsymbol{\Psi}_{\text{s}}}\frac{1}{(C^l)^2}\sum_{i=1}^{C^l}\sum_{j=1}^{C^l}(\mathcal{G}^l_{\text{cs},ij}-\mathcal{G}^l_{\text{s},ij})^2,\quad(12)$$

式中：C^l 表示 l 层上的特征通道数；$\mathcal{G}^l_{\text{cs},ij}=\mathcal{F}^l_{\text{cs},i}\cdot\mathcal{F}^l_{\text{cs},j}$；$\mathcal{F}^l_{\text{cs},i}$ 表示特征图 $\mathcal{F}^l_{\text{cs}}$ 的第 i 个响应层。如表 1，2 最后一列所示，各方法在新风格度量指标上的定量计算结果大致遵循其在原有指标上的表现，即本文方法风格差异最小，STTR[18]风格损失值最大，进一步佐证了本文方法的有效性。

3.4 消融实验

本文通过开展消融实验对方法不同组件进行有效性验证，具体包括风格参数提取器、非线性投影、CNN 解码器以及恒等损失，相应风格化结果如图 6 后四列所示。此外，本文基于 3.3 节中构建的内容-风格图像对和三项评价指标定量对比了不同消融设置下的方法。如表 3 所示，完整方法两项差异均最小，可见本文方法各组成部分缺一不可。此外，CNN 解码器与恒等损失对于实施准确稳定的风格化，其影响程度较大。

3.4.1 风格参数提取器

为验证 Transformer 风格参数提取器的表征能力以及风格参数的有效性，本文直接对风格编码 $\boldsymbol{\Phi}_{\text{s}}$ 进行通道降维，以作为 Transformer 图像解码器中所使用的风格参数 $\hat{\boldsymbol{\mathcal{S}}}$，即移除 Transformer 风格参数提取器，作为消融(a)方法。如图 6(a)所示，消融(a)方法的风格化结果仍能有效保留内容语义并呈现特定风格，但整体纹理感不及完整方法，特别是图像背景样式被削弱(如行 1、行 3)，与 StyTr^2[16]效果相似。然而，这也侧面印证了将 LN 替换为 CIN 的可行性以及通过特征配准对齐实现图像风格迁移的有效性。

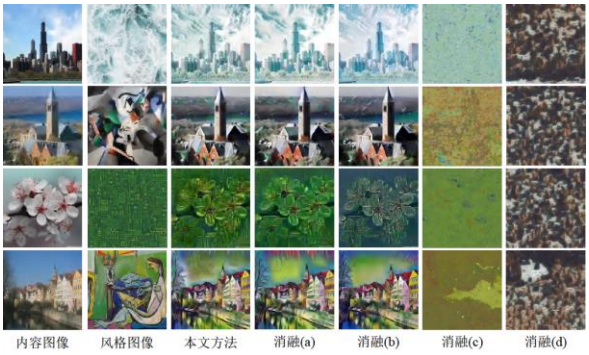

图 6 不同消融设置下的风格迁移效果对比

Fig.6 Comparison of style transfer results under different ablation settings

表 3 不同消融设置下的平均风格化损失对比

Table 3 Comparison of average stylization loss under different ablation settings

方法	$\mathscr{L}_{\text{con}}\downarrow$	$\mathscr{L}_{\text{sty}}-\mu,\sigma\downarrow$	$\mathscr{L}_{\text{sty}}-$ *Gram*矩阵↓
消融(a)	0.70	1.02	3.15
消融(b)	0.64	1.37	4.71
消融(c)	1.88	11.36	18.61
消融(d)	1.93	6.75	16.05
完整方法	0.69	0.94	3.02

3.4.2 非线性投影

本文采用预训练 VGG19 网络进行图像块划分与编码，以中和视觉 Transformer 的形状偏向。为进行有效性验证，本文将其替换为原简单线性投影方式，作为消融(b)方法，即对输入图像进行一次卷积操作，将卷积核大小、卷积步长和输出通道数分别设置为 m、m、C，从而得到图像块特征序列 $\boldsymbol{\mathcal{E}}_{\text{c}}$ 和 $\boldsymbol{\mathcal{E}}_{\text{s}}$。如图 6(b)所示，消融(b)方法的

风格化结果无论是色彩分布还是纹理样式均偏离参考风格(如行 1、行 3)，可见非线性投影能够提升模型对纹理色彩的学习能力，更加契合风格迁移任务目标。

3.4.3 CNN 解码器

如文献[11]中所述，直接将 Transformer 解码输出上采样至图像空间，会产生严重的棋盘效应，使图像块间的样式区分度较低。本文将 CNN 解码器替换为由全连接层、ReLU 激活与上采样层构成的多层感知机，作为消融(c)方法，其风格化结果如图 6(c)所示，验证了 CNN 解码器相对于直接上采样的优越性。

3.4.4 恒等损失

本文方法在训练阶段共包含四项分支损失，其中 $\mathcal{L}_{con}$ 和 $\mathcal{L}_{sty}$ 为图像风格迁移任务中的必要损失，分别用于保留内容与融合样式。本文移除两个恒等损失项 $\mathcal{L}_{id}^{1}$ 和 $\mathcal{L}_{id}^{2}$，作为消融(d)方法，以验证其对风格化效果的影响，如图 6(d)所示，当移除恒等损失时，消融(d)方法的迁移图像存在内容失真与风格漂移问题，模型无法学习到施加纹理样式的合理位置，引入大量冗余变动而导致内容语义紊乱。

4 结论

1) 本文结合条件实例归一化在任意风格迁移中的成功经验，并借助视觉 Transformer 强大的表达能力，提出了一种基于双路视觉 Transformer 的任意图像风格迁移方法 *Bi-Trans*，相较于基准方法增强图像块编码的纹理偏向，更加准确地塑造目标风格样式的色彩分布与显著纹理，解决了现有基于 CNN 的图像风格迁移方法中存在的感受野有限、混淆图像域及纹理偏向等问题，兼具较高的内容保真度与风格还原度。

2) 实验结果表明了本文方法相较于现有代表性方法的先进性与优越性，验证了本文方法各组成部分对于实现高质量图像风格迁移的不可或缺性与有效性。

3) 尽管本文方法堆叠较少的网络层即可有效获得全局感受野，但相较于基于 CNN 的 IST 方法存在计算复杂度高等问题，其在 Transformer 内容图像编码阶段、Transformer 风格图像编码阶段以及 Transformer 图像解码阶段均涉及到两两图像块间的相关性计算，在图像渲染效率上有待进一步提升。

参考文献（References）

[1] ZHANG Y,HUANG N,TANG F,et al.Inversion-based style transfer with diffusion models[C]//IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos,CA:IEEE Computer Society,2023:10146-10156.

[2] ZHANG Y,DONG W,TANG F,et al.ProSpect: prompt spectrum for attribute-aware personalization of diffusion models[J].ACM Transactions on Graphics,2023,42(6):1-14.

[3] WANG Z,ZHAO L,XING W.Stylediffusion: controllable disentangled style transfer via diffusion models[C]//IEEE International Conference on Computer Vision. Los Alamitos,CA:IEEE Computer Society,2023: 7677-7689.

[4] ROBIN R,ANDREAS B,DOMINIK L,et al.High-resolution image synthesis with latent diffusion models.[C]//IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos,CA:IEEE Computer Society,2022:10684-10695.

[5] JING Y,YANG Y,FENG Z,et al.Neural style transfer: a review[J].IEEE Transactions on Visualization and Computer Graphics,2019,26(11):3365-3385.

[6] GATYS L A,ECKER A S,BETHGE M.Image style transfer using convolutional neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos,CA:IEEE Computer Society,2016:2414-2423.

[7] JOHNSON J,ALAHI A,LI F F.Perceptual losses for realtime style transfer and super-resolution[C]//European Conference on Computer Vision. Berlin Germany:Springer-verlag,2016，9906:694-711.

[8] ULYANOV D,LEBEDEV V,VEDALDI A, et al.Texture networks: feed-forward synthesis of textures and stylized images[C]//International Conference on Machine Learning. San Diego,GA:JMLR,2016，48:1349-1357.

[9] LIN T,MA Z,LI F,et al.Drafting and revision: laplacian pyramid network for fast high-quality artistic style transfer[C]//IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos,CA:IEEE Computer Society,2021:5141-5150.

[10] GHIASI G,LEE H,KUDLUR M,et al.Exploring the structure of a real-time arbitrary neural artistic stylization network[C]//British Machine Vision Conference. Great Britain:BMVA,2017:1-27.

[11] HUANG X,BELONGIE S.Arbitrary Style transfer in realtime with adaptive instance normalization[C]//IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos,CA:IEEE Computer Society,2017:1501-1510.

[12] LI Y,FANG C,YANG J,et al.Universal style transfer via feature transforms[C]//Annual Conference on Neural Information Processing Systems. La Jolla, CA:NIPS,2017，30:1-11.

[13] DAE Y P,KWANG H L.Arbitrary style transfer with style-attentional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos,CA:IEEE Computer Society,2019:5880-5888.

[14] LIU S,LIN T,HE D, et al.Adaattn: revisit attention mechanism in arbitrary neural style transfer[C]//IEEE International Conference on Computer Vision. Los Alamitos,CA:IEEE Computer Society,2021:6649-6658.

[15] Chandran P,Zoss G,Gotardo P,et al.Adaptive convolutions for

structure-aware style transfer[C]//IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos,CA:IEEE Computer Society,2021:7972-7981.

[16] DENG Y,TANG F,DONG W,et al.StyTr^2: image style transfer with transformers[C]//IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos,CA:IEEE Computer Society,2022:11326-11336.

[17] ZHANG C,YANG J,WANG L,et al.S2WAT: image style transfer via hierarchical vision transformer using Strips Window Attention[J/OL].arXiv preprint, 2022[2023-06-19]. https://arxiv.org/abs/2210.12381

[18] WANG J,YANG H,FU J,et al.Fine-grained image style transfer with visual transformers[C]//Asian Conference on Computer Vision. Berlin Germany:Springer-verlag,2022:841-857.

[19] ZHANG C,DAI Z,CAO P,et al.Edge enhanced image style transfer via transformers[C]//ACM International Conference on Multimedia Retrieval. New York, NY: ACM,2023:105-114.

[20] FENG J,ZHANG G,LI X, et al.A compositional transformer based autoencoder for image style transfer[J].Electronics,2023,12(5):1184.

[21] LECUN Y,BOTTOU L,BENGIO Y,et al.Gradient-based learning applied to document recognition.[J].Proceedings of the IEEE,1998,86(11):2278-2324.

[22] KRIZHEVSKY A,SUTSKEVER I,HINTON G E.Imagenet classification with deep convolutional neural networks. [J].Communications of the ACM,2017,60(6): 84-90.

[23] GEIHOS R,RUBISCH P,MICHALIS C,et al.ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness[C]//International Conference on Learning Representations. Washington DC: ICLR,2019:1-22.

[24] WEI H,DENG Y,TANG F,et al.A comparative study of CNN- and Transformer-based visual style transfer[J].Journal of Computer Science and Technology,2022,37(3):601-614.

[25] DOSOVITSKIY A,BEYER L,KOLESNIKOV A,et al.An image is worth 16x16 words: Transformers for image recognition at scale.[J/OL].arXiv preprint,2020[2023-06-19]. https://arxiv.org/abs/2010.11929.

[26] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[J].Advances in Neural Information Processing Systems. La Jolla, CA:NIPS,2017,30:5998-6008.

[27] LIN T Y,MAIRE M,BELONGIE S J,et al.Microsoft COCO: common objects in context[C]//European Conference on Computer Vision. Berlin Germany:Springer-verlag,2014 ，8693:740-755.

[28] FRED P,BRANDY M.Wiki Art Gallery, Inc.: a case for critical thinking[J].Issues in Accounting Education,2011,26(3):593-608.

Dual-channel Vision Transformer Based Image Style Transfer

JI Zongxing，BEI Jia*, LIU Runze, REN Tongwei

（State Key Laboratory for Novel Software Technology , Nanjing University, Nanjing 210093, China）

Abstract: Image style transfer (IST) aims to adjust the visual properties of a content image in accordance with a style image to generate a stylized image with visual appeal, which preserves the original content and appears specific style patterns. Most existing methods neglect the encoding discrepancies between different image domains, focusing on extracting local image features while ignoring the importance of global contextual information. To address this issue, a novel image style transfer method based on dual-channel vision transformer is proposed, called *Bi-Trans*, which constructs dedicated encoders for different image domains and extracts parameter vectors to discretely represent the reference style, calibrating the content image to the target style domain through cross-attention mechanism and conditional instance normalization. Experimental results demonstrate that the proposed method is superior to state-of-the-arts in terms of both content retention and style restoration.

Keywords: Image Style Transfer; Vision Transformer; Arbitrary Stylization; Conditional Instance Normalization; Attention Mechanism