基于ViT语义指导与结构感知增强的艺术风格迁移

潘书煜[1)] 赵征鹏[1),2)] 阳秋霞[1)] 普园媛[1),3)] 谷金晶[1)] 徐 丹[1)]

[1)](云南大学信息学院 昆明 650500)

[2)](云南省微光夜视探测及智能视觉导航重点实验室 昆明 650500)

[3)](云南省高校物联网技术及应用重点实验室 昆明 650500)

摘 要 艺术风格迁移是计算机视觉领域一个长期的研究热点,该任务旨在将参考风格图像的艺术风格迁移到内容图像中,同时保持内容图像的语义结构不变。目前基于深度学习的艺术风格迁移方法依然面临一项主要挑战:现有方法在迁移过程中无法很好地保持内容域到风格域的语义结构跨域一致性,从而导致风格化结果的内容保真度低、风格化不一致。针对以上问题,本文提出了一种基于ViT(Vision Transformer)语义指导与结构感知增强的艺术风格迁移方法。首先,利用预训练的DINO-ViT模型在内容域和风格域建立强大且一致的内容结构表示,并设计了两种损失函数:(1)DINO keys自相似性的语义结构损失,以保持内容源的跨域一致性;(2)DINO特征空间的知识蒸馏损失,以提升编码器的特征提取能力。其次,为进一步增强模型的结构感知能力,提出了基于拉普拉斯算子的空间结构损失和基于小波变换的频域纹理损失,从空间域和频率域两方面增强了对边缘轮廓与细致纹理的约束。在通用数据集MS COCO和WikiArt上的定性与定量结果表明,本文方法不仅可以产生内容保真度高、风格化一致的结果,还能推广应用于现有方法以进一步改善生成结果的视觉质量。其中,与基线方法CAP-VST相比,本文方法的SSIM值提升0.079,CLIP-IQA值提升0.024,LPIPS值小0.096,Content Loss值小1.035;将本文方法应用于其他现有方法后,SSIM值最优提升0.135,CLIP-IQA值最优提升0.011,LPIPS值最优小0.108,Content Loss值最优小1.244,证明了本文方法在艺术风格迁移任务中的有效性与灵活性。

关键词 艺术风格迁移;Vision Transformer;知识蒸馏;结构感知;拉普拉斯算子;小波变换

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2025.02131

Structure-Enhanced Artistic Style Transfer via ViT Semantic Guidance and Structure-Aware Enhancement

PAN Shu-Yu[1)] ZHAO Zheng-Peng[1),2)] YANG Qiu-Xia[1)]

PU Yuan-Yuan[1),3)] GU Jin-Jing[1)] XU Dan[1)]

[1)](School of Information Science and Engineering, Yunnan University, Kunming 650500)

[2)](Yunnan Key Laboratory of low-light Night Vision Technology and Intelligent Visual Navigation, Kunming 650500)

[3)](University Key Laboratory of Internet of Things Technology and Application in Yunnan Province, Kunming 650500)

Abstract Artistic style transfer has long been a prominent research topic in the field of computer vision, aiming to transfer the artistic style of a reference image onto a content image while preserving the semantic structure of the content. Despite significant progress, deep learning-based artistic style transfer methods still face a major challenge: existing approaches struggle to maintain

收稿日期:2024-08-28;在线发布日期:2025-05-22。本课题得到国家自然科学基金地区项目(62362070,61761046)、国家自然科学基金面上项目(61271361)、云南省科技厅科技计划基础研究专项-重点项目(202401AS070149)、云南省应用基础研究计划资助项目重点项目(202001BB050043)、云南省微光夜视探测及智能视觉导航重点实验室(202449CE340004)、云南大学第三届专业学位研究生实践创新项目(ZC-23235984)、云南大学第十五届研究生科研创新项目(KC-23235986)资助。**潘书煜**,硕士研究生,主要研究领域为计算机视觉、图像风格迁移。E-mail:panshyu@stu.ynu.edu.cn。**赵征鹏**,硕士,教授,主要研究领域为信号与信息处理、计算机系统及应用。**阳秋霞**,博士研究生,主要研究领域为计算机视觉、图像翻译。**普园媛(通信作者)**,博士,教授,主要研究领域为数字图像处理、视觉艺术科学理解。E-mail:yuanyuanpu@ynu.edu.cn。**谷金晶**,博士,讲师,主要研究领域为跨媒体语义分析与理解、视频异常行为识别。**徐 丹**,博士,教授,主要研究领域为图形绘制技术、图像融合。

cross-domain semantic structural consistency between the content and style domains during the transfer process. To address this challenge, early methods utilized deep features extracted by pre-trained convolutional neural networks (such as VGG-19) as content representations and employed perceptual losses to preserve content structure. However, these pre-trained models inherently exhibit limitations in modeling content features, inevitably leading to the loss of content information and degradation of structural integrity. More recent approaches have introduced invertible neural flow models as encoders/decoders to better maintain content affinity and mitigate content leakage. Although some progress has been achieved, these methods are still constrained by the lack of explicit semantic and structural guidance, making it difficult to ensure cross-domain consistency of the content source, which continues to result in stylized outputs with low content fidelity and inconsistent stylization. To address the above issues, this paper proposes an artistic style transfer method based on ViT (Vision Transformer)-guided semantic supervision and structure-aware enhancement. First, we explore and analyze the differences in feature spaces among mainstream pre-trained ViT models (original ViT, CLIP, and DINO), demonstrating that the self-supervised DINO ViT model exhibits superior scene understanding and semantic localization capabilities, whereas supervised models such as CLIP and the original ViT are relatively deficient. This provides new insights into the selection and use of pre-trained ViT models for artistic style transfer tasks. Based on this analysis, we leverage a pre-trained DINO-ViT model to establish strong and consistent content structure representations across the content and style domains. Two loss functions are specifically designed: (1) a semantic structure loss based on the self-similarity of DINO keys, aiming to preserve the cross-domain consistency of the content source; and (2) a knowledge distillation loss in the DINO feature space, designed to enhance the encoder's feature extraction capabilities. Furthermore, to further strengthen the model's structure awareness, we propose a spatial structure loss based on the Laplacian operator and a frequency-domain texture loss based on wavelet transform, reinforcing constraints on edge contours and fine textures from both the spatial and frequency domains. Qualitative and quantitative results on the MS COCO and WikiArt datasets demonstrate that the proposed method not only produces stylized outputs with high content fidelity and consistent stylization but can also be flexibly applied to existing methods to further improve the visual quality of generated results. Compared to the baseline method CAP-VST, our method achieves an improvement of 0.079 in SSIM, 0.024 in CLIP-IQA, a reduction of 0.096 in LPIPS, and a reduction of 1.035 in Content Loss. Moreover, when applied to other existing methods, our approach achieves optimal improvements of 0.135 in SSIM, 0.011 in CLIP-IQA, a reduction of 0.108 in LPIPS, and a reduction of 1.244 in Content Loss, demonstrating its effectiveness and flexibility in the task of artistic style transfer.

Keywords artistic style transfer; Vision Transformer; knowledge distillation; structure-awareness; Laplacian operator; wavelet transform

1 引 言

艺术风格迁移是计算机视觉领域一个长期的研究热点,该任务旨在将艺术风格从参考图像迁移到内容图像,例如将文森特·梵高的《星月夜》的风格迁移到日常照片中。自从 Gatys 等[1]首次提出了一种利用预训练的深度卷积神经网络(Deep Con-volutional Neural Network,DCNN)来分离和重组任意图像的内容和风格后,艺术风格迁移领域迎来了前所未有的蓬勃发展[2-7]。

长久以来,艺术风格迁移面临的一项主要挑战是在迁移过程中保持内容域到风格域的语义结构跨域一致性,其关键在于模型对内容图像的内容语义

与空间结构的感知能力:模型的感知能力越强,跨域一致性保持越好,从而生成内容保真度更高、风格化更一致的结果图像,如图 1(b)和(e)所示;反之,则会生成视觉质量较低的图像,如图 1(c)和(f)所示。为此,现有的方法采取了许多措施来提升模型的语义与结构感知能力。早期的 AdaIN[3] 和 WCT[4] 将预训练的 DCNN(如 VGG-19[8])提取到的深度特征作为图像的内容表示,并结合感知损失[2] 以维持内容结构的不变。然而,由于 VGG-19 本身是为捕捉对象级信息的分类任务所设计,对内容特征的建模存在固有缺陷[9],因此当其在风格迁移任务中使用时会不可避免地导致内容信息丢失,破坏内容结构。最近,ArtFlow[10] 与 CAP-VST[9] 引入了可逆神经流模型[11] 作为编码器/解码器以替换 VGG-19,试图保持内容亲和度,解决内容泄漏问题。尽管取得了一定进展,但受限于缺乏显式的语义与结构指导,此类方法无法保持内容源的跨域一致性,生成的风格化图像依然存在内容保真度低、风格化不一致的问题。

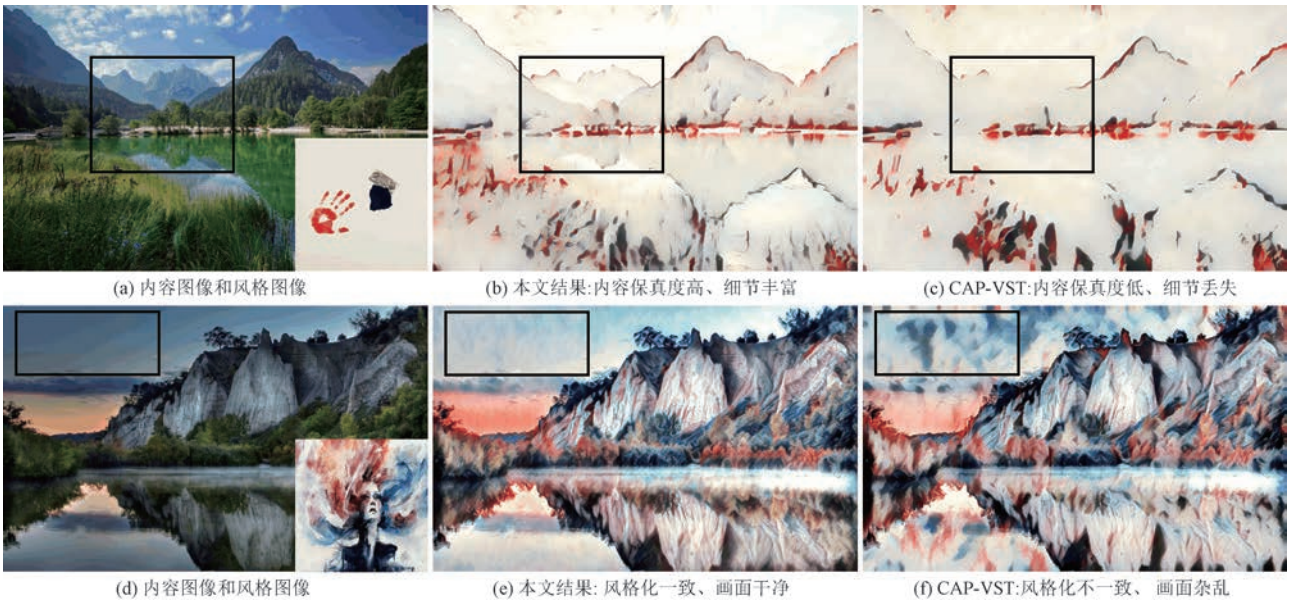


图 1　本文方法与现有方法(CAP-VST[9])的风格化结果对比示意图

面对以上挑战,我们注意到在图像翻译[12] 领域最近的发展中,借助 Vision Transformer (ViT)[13] 等预训练大模型的能力可以对生成进行更细粒度的控制。受到上述观察的启发,我们进行了如下思考:能否通过引入外部预训练大模型的语义指导提升模型的结构感知能力,从而保持内容语义的跨域一致性?

为了实现这个目标,本文引入了 DINO-ViT[14](下称 DINO)——一种以自监督方式预训练的 ViT 模型。关于 DINO,已有研究证明[14] 其可以学习强大且有意义的视觉表征,能够应用于图像翻译、图像检索、实例分割等多种下游任务。对此,本文首先研究了 DINO 的特征空间,证明了其能跨内容域和风格域同时捕获稳健且一致的结构语义信息,优于另外一种主流的预训练 ViT 模型 CLIP[15]。基于此发现,本文提出了一种基于 ViT 语义指导的艺术风格迁移方法,能够有效保持内容语义的跨域一致性。首先,本文计算了一种语义结构损失。该损失利用 DINO 提取图像的深层 ViT keys 特征,并将 keys 特征之间的自相似性作为图像的内容结构表示,以约束内容图像与风格化图像之间的结构相似度,显著缓解生成结果内容保真度低、风格化不一致的问题。此外,为了充分利用 DINO 强大的语义表征能力,本文还提出了一种知识蒸馏损失,通过模仿 DINO 的特征空间,提炼丰富且多样的大模型知识,有效增强了编码器的特征提取能力,改善了生成质量。

同时,为了进一步提升模型的结构感知能力,确保空间连贯性,本文还引入了显式的外部结构先验。具体而言,本文首先通过拉普拉斯算子[16-17] 滤波输入图像以获得边缘表示,该表示包含了具有层次感的空间结构信息——如对象轮廓、场景布局及景深关系。然后通过 L1 损失规范内容图像与风格化图像之间的边缘表示,为风格迁移模型提供显式结构引导,增强全局内容的保真度。此外,受到最近神经网络更偏向于拟合低频信号而容易忽略高频信号的发现[18-21] 的启发,本文进一步利用了小波变换[22] 生成的频域表示来鼓励模型拟合高频纹理信息。

本文在风格迁移任务的通用数据集 MS COCO[23] 和 WikiArt[24] 上评估了所提出方法的有效

性。结果表明,相较现有方法,本文方法显著提升了内容结构保真度和风格化一致性。此外,本文方法与现有的模型是可兼容的,将本文方法集成到现有模型中可获得进一步的性能提升。其中,与基线方法 CAP-VST 相比,本文方法的 SSIM 值提升 0.079,CLIP-IQA 值提升 0.024,LPIPS 值小 0.096,Content Loss 值小 1.035;将本文方法应用于其他现有方法后,SSIM 值最优提升 0.135,CLIP-IQA 值最优提升 0.011,LPIPS 值最优小 0.108,Content Loss 值最优小 1.244,证明了本文方法在艺术风格迁移任务中的有效性与灵活性。

2 相关工作

2.1 艺术风格迁移

艺术风格迁移旨在生成具有内容图像的结构和参考图像的风格的图像。Gatys 等[1]率先提出了基于迭代优化的神经风格迁移(Neural Style Transfer,NST)。为了加速推理,Johnson 等[2]将迭代优化过程近似为前馈网络,并通过单次快速前向传播实现了实时风格迁移。为了提升泛化性,AdaIN[3]与 WCT[4]被提出以实现任意风格迁移。其中,AdaIN[3,25]将内容特征的均值和标准差替换为风格特征的均值和标准差。WCT[4]利用奇异值分解对图像进行白化后再着色。之后,SANet[26]引入了注意力机制,利用内容特征与风格特征之间的空间相关性重新排列风格特征。AdaAttN[27]结合了 AdaIN[3]和 SANet[26],同时利用全局方法和局部方法的优势。遵循同样的思路,TSSAT[28]先后在全局和局部维度使用了 AdaIN[3]并加以整合。

近些年陆续出现了许多思路新颖的方法。IECST[29]、CCPL[30]和 MicroAST[31]引入了对比学习[32]来增强特征表示能力,提高生成质量。ArtFlow[10]和 CAP-VST[9]通过神经流模型[11]提升内容亲和度,解决基于 VGG 框架[8]的算法产生的内容泄漏问题。尽管取得了一定成效,但由于缺乏更加准确的显式结构和语义表示,这些方法仍然无法保持内容源的跨域一致性,存在内容保真度低和风格化不一致的问题,如图 1 (c)和(f)所示。与现有的方法不同,本文引入了预训练大模型 DINO-ViT[14],并结合拉普拉斯算子[16]和小波变换[22],提出了一种 DINO 语义指导的结构感知增强的训练方法,显著缓解了上述问题,如图 1 (b)和(e)所示。

2.2 ViT 特征表示

自从问世以来,ViT 架构[13]已经在图像分类任务中展现出与最先进的 CNN(Convolutional Neural Network)架构同样具有竞争力的表现,并且表现出很强的鲁棒性。其中,DINO-ViT 是一种使用自蒸馏方法进行训练的无标签 ViT 模型[14],它学习到的特征表示已在图像检索和图像分割等多个下游任务中证明了有效性。

最初,Amir 等[33]证明了 DINO-ViT 特征作为密集视觉描述算子的能力,即深度 DINO 特征可以在细粒度的空间尺度上捕获诸如语义对象在内的丰富语义信息。同时,他们还观察到不同但相关的对象类之间会共享同样的特征表示,证明了 DINO 鲁棒的语义建模能力。

此外,在图像翻译领域,Tumanyan 等[12]进一步利用 DINO 特征解耦并拼接图像的结构表示与外观表示,实现了自然图像之间的语义外观风格迁移,展现了其在结构语义跨域一致性方面的优势。最近,Zhou 等[34]通过 DINO 对人脸域的面部风格化进行了探索,展示了在将自然图像转换为艺术图像的任务中使用 DINO 的潜力。受到以上工作的启发,本文在艺术风格迁移任务中引入了 DINO-ViT。

2.3 知识蒸馏

知识蒸馏是一种通用算法,通过迁移预训练教师网络的知识来监督学生网络的训练。知识蒸馏最初是通过模仿模型集合的输出为模型压缩所设计[35]。在此之后,Jimmy 等[36]通过模仿 logits 进一步将深层网络压缩为更浅但更宽的网络。Hinton 等[37]提出了一种更通用的知识蒸馏算法,即应用教师模型的预测作为软标签。最近 Cui 等[38]利用知识蒸馏解决了小样本生成任务中 GAN 判别器的过拟合问题。目前,知识蒸馏在各种任务中都得到了广泛应用,例如图像分类、域自适应、目标检测以及风格迁移。

对于风格迁移任务,与 Wang 等[39]利用知识蒸馏压缩 VGG 编码器-解码器对以生成轻量化风格迁移模型不同,本文将其用于提炼预训练大模型 DINO 中丰富且多样的知识,使编码器模仿 DINO 的特征空间,旨在增强特征提取的能力。

2.4 计算机视觉中的小波变换

Schwarz 等[19-20]证明了深度神经网络在训练和推理过程中存在频域偏差,即模型倾向于捕捉低频信号而忽略高频信息,进而导致生成结果出现伪影。小波变换作为一种经典信号处理方法,由于其良好的收敛性和对任意信号的紧凑表示能力[40-41],已广

泛应用于图像生成[17,21,42-43]、图像超分[44-45]、图像翻译[46]、姿态迁移[47]、风格迁移[48]等任务以解决频域偏差问题。在图像生成的小样本生成任务中,Yang 等[17,21,43]将小波变换与 GAN 网络结合,设计了基于小波变换的高频判别器来增强生成器的频率意识,鼓励模型区分不同频域信号。Phung 等[42]将小波变换与扩散模型结合,提出了一种基于小波的扩散方案,在改善生成质量的同时保持了实时性能。在图像超分任务中,Li 等[44]利用小波变换能够在不同尺度上表示特征上下文和纹理信息的优势提出了一种基于小波的纹理重构网络。在图像翻译任务中,Kim 等[46]利用小波变换的多尺度特性提出了一种小波域高频损失,通过多层级小波分解使模型更加关注高频信息,以克服现有方法偏向低频的局限性。Ma 等[47]设计了一种基于小波感知的人像姿态迁移模型,在小波域中融合注意力与光流特征以发挥二者在不同频域的优势。

在真实感风格迁移任务中,Yoo 等[48]提出了一种基于 WCT 的小波校正风格迁移方法。该方法将小波变换嵌入深度网络,通过高频残差跳连接缓解模型的频率偏差,使特征在风格化过程中同时保留结构信息和 VGG 特征空间中的统计分布,以改善空间变形和不真实伪影问题。受到上述方法的启发,本文在艺术风格迁移任务中引入了小波变换,并设计了频域纹理与重构损失,以帮助模型更好地生成高频信号,增强纹理细节。

2.5　图像生成中的结构表示

稳健的结构表示对图像生成任务至关重要,因为它决定了模型是否能够保持输入图像的原始质量[49]。因此,各种图像生成方法都致力于引入结构表示以改善生成效果。在风格迁移任务中,Liu 等[50]基于深度边缘检测模型 HED[51]提出了一种精细化网络,使风格化图像能够在不同层级上保留细节信息。类似地,Wu 等[52]提出了一种边缘增强模块,通过 HED 和平均池化层多尺度、多层级地捕捉关键边缘信息,以增强生成图像的结构一致性。在图像卡通化任务中,Gao 等[53]设计了改进版 Sobel 算子并提出了一种纹理显著性自适应注意力机制,用于自适应地从每个输入图像 mini-batch 中采样具有最显著边缘的局部图像块。Cho 等[49]在扩散模型中引入了一种基于 1×1 卷积的结构保持网络,用于控制输入图像的内容与风格保留程度。

近期,在小样本图像生成任务中,Yang 等[17]将拉普拉斯算子[16]与 GAN 网络结合设计了一种全局结构判别器,通过提取包含丰富全局结构信息的拉普拉斯表示显式地引导模型生成具有合理布局与轮廓的图像。类似的,本文在艺术风格迁移任务中引入拉普拉斯算子并构建了空间结构损失与空间重构损失,以改善生成结果的内容保真度。

3　本文方法

3.1　方法概述

令 I_C 和 I_S 分别为内容照片和真实艺术作品。本文目标是训练一种风格迁移模型,能够将任意艺术作品 I_S 的风格转移到内容图像 I_C 上。本文方法建立在艺术风格迁移模型 CAP-VST[9]上,该模型由编码器 E、转换模块 cWCT 和解码器 D 组成。具体来说,E 和 D 由同一个可逆残差神经网络[10-11]组成。该网络的前向过程用于提取图像特征,反向过程用于解码风格化特征以生成图像。cWCT 是一个基于 Cholesky 分解的白化着色变换模块[4],能够利用矩阵特征值分解灵活地将全局风格特征匹配到内容特征上。本文基于 ViT 语义指导与结构感知增强策略提出了一种损失函数训练方法,旨在改进上述主干网络 CAP-VST 的生成质量,具体方法概述如下:如图 2 所示,本文方法总体流程由三部分组成。

(1)风格化路径。该路径为风格迁移的主过程。在训练阶段,艺术风格迁移模型首先将输入的内容图像 I_C 与风格图像 I_S 合成为风格化结果 I_{CS} 输出。然后,利用输出图像与输入图像计算训练损失函数并优化模型参数。该路径中将计算 5 种损失函数:风格损失 L_{style}、抠图拉普拉斯损失 L_{mat}、ViT 语义结构损失 L_{vit}^{str}、空间结构损失 L_{lap}^{str}、频域纹理损失 L_{fre}^{tex}。其中,L_{style} 与 L_{mat} 为主干网络 CAP-VST 的损失函数,L_{vit}^{str}、L_{lap}^{str}、L_{fre}^{tex} 为本文基于 DINO-ViT[14]、拉普拉斯算子[16]、小波变换[22]提出的损失函数。训练完成后,推理阶段仅保留训练好的艺术风格迁移模型以进行图像风格化任务。

(2)重构路径。该路径为风格迁移的辅助训练任务,旨在提升模型的内容感知与保持能力。在该路径中,本文将风格化图像 I_{CS} 和内容图像 I_C 分别作为内容表示和风格表示,输入风格迁移模型并鼓励其将内容图像 I_C 的风格信息迁移至风格化图像 I_{CS},进而重构内容图像 $\hat{I}_C$。相应的,本文采用了 3 种重构损失函数:循环一致性像素重构损失 L_{pix}^{rec}[54]、空间重构损失 L_{lap}^{rec}、频域重构损失 L_{fre}^{rec}。其中,

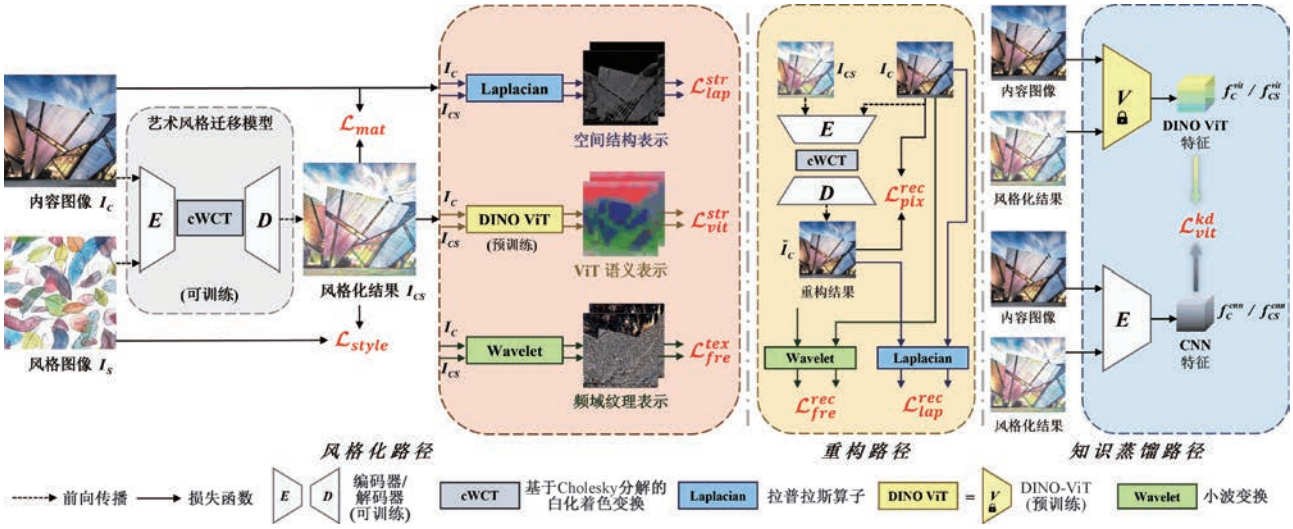


图 2 本文方法的总体流程(分为三个部分:(1)风格化路径。艺术风格迁移模型首先将输入的内容图像 I_C 与风格图像 I_S 合成为风格化结果 I_{CS} 输出,然后利用输出图像与输入图像计算训练损失函数以优化模型参数。其中,本文基于 DINO-ViT 模型[14]、拉普拉斯算子[16]、小波变换[22]提出了 3 种损失函数:ViT 语义结构损失 L_{vit}^{str}、空间结构损失 L_{lap}^{str}、频域纹理损失 L_{fre}^{tex}。(2)重构路径。该路径将风格化图像 I_{CS} 和内容图像 I_C 分别作为内容表示和风格表示,输入风格迁移模型并鼓励其将内容图像 I_C 的风格信息迁移至风格化图像 I_{CS},进而重构内容图像 $\hat{I}_C$。其中,本文基于拉普拉斯算子[16]与小波变换[22]设计了 2 种重构损失函数:空间重构损失 L_{lap}^{rec}、频域重构损失 L_{fre}^{rec},它们是风格化路径中空间结构损失 L_{lap}^{str} 和频域纹理损失 L_{fre}^{tex} 的变体(仅输入图像略有不同),旨在从空间域与频率域两方面提升模型对内容特征与风格化特征的建模鲁棒性。(3)知识蒸馏路径。该路径将内容图像 I_C 与风格化结果 I_{CS} 同时输入可训练的主干网络编码器 E 与预训练 ViT 模型 V 中,获得 CNN 特征 f_C^{cnn}/f_{CS}^{cnn} 与 DINO ViT 特征 f_C^{vit}/f_{CS}^{vit}。然后,利用 ViT 知识蒸馏损失 L_{vit}^{kd} 约束 CNN 特征与 ViT 特征,迫使主干网络编码器 E 学习 DINO 的编码空间以加强特征提取性能。更多细节请参考第 3 节。)

L_{pix}^{rec} 为现有方法常用的重构损失,旨在从像素层面约束重构图像与原始内容图像的一致性。L_{lap}^{rec} 与 L_{fre}^{rec} 是本文基于拉普拉斯算子[16]和小波变换[22]设计的两种重构损失函数,它们是风格化路径中空间结构损失 L_{lap}^{str} 和频域纹理损失 L_{fre}^{tex} 的变体(仅输入图像略有不同),旨在从空间域与频率域两方面提升模型对内容特征与风格化特征的建模鲁棒性。

(3)知识蒸馏路径。该路径为风格迁移的辅助训练任务,包含一种基于 DINO-ViT[14]设计的知识蒸馏损失,旨在利用预训练 ViT 模型 V 提升主干网络编码器 E 的特征提取能力。具体而言,在训练阶段本文将内容图像 I_C 与风格化结果 I_{CS} 同时输入可训练的主干网络编码器 E 与预训练 ViT 模型 V 中,获得 CNN 特征 f_C^{cnn}/f_{CS}^{cnn} 与 DINO ViT 特征 f_C^{vit}/f_{CS}^{vit}。然后,利用 ViT 知识蒸馏损失 L_{vit}^{kd} 约束 CNN 特征与 ViT 特征,迫使主干网络编码器 E 学习 DINO 的编码空间以加强特征提取性能。

3.2 基于 ViT 语义指导的损失

3.2.1 ViT 语义指导分析

自从 ViT 问世以来,在自然语言处理和计算机视觉任务中,最主流的预训练 ViT 大模型是 CLIP[15],这是一种通过文本-图像对训练的弱监督文本-图像嵌入模型。由于其在连接文本域和图像域方面表现出色,目前已广泛应用于文本引导的零样本风格迁移[55-56]任务中。然而,最近 Wysoczańska 等[57-58]指出,CLIP 缺乏空间意识,不适合密集的计算机视觉任务。相反,最近的自监督方法能够产生定位空间对象属性的强视觉表示,其中以具有语义意识的仅使用图像训练的 DINO[14]最佳。此外,Zhou 等[34]证明了 DINO 比 CLIP 更适合人脸风格化任务。

鉴于以上先验,本文期望 DINO 也能在艺术风格迁移中跨内容域与风格域捕获一致的鲁棒结构表示。因此,本文按照 Zhou 等[12,34]的方式可视化 DINO 与 CLIP 的中间特征(keys 与 tokens),并分析

它们在艺术风格迁移任务中的表现。注意,由于[CLS] tokens 在[12]中表示风格外观特征,因此其不包含在可视化的 tokens 内。如图 3 所示,本文首先将真实域的内容图像、艺术域的风格图像和风格化之后的结果图像输入到 CLIP 与 DINO 中,然后通过主成分分析(Principal Component Analysis,PCA)[59]可视化来自不同层的 keys 和 tokens。为公平起见,CLIP 与 DINO 均选用 ViT-B/16 架构(共 12 层,每层的嵌入维度为 768,多头注意力机制的 head 数为 12),同时选用第 3、6、12 层作为低、中、高层表示,降维后的主成分个数为 3。

如图 3 所示,从 PCA 结果中可以看出,虽然 CLIP 能够大致描述前景对象的边界轮廓,但背景中存在杂乱的纹理伪影,这会影响风格迁移模型在训练过程中对目标语义结构的学习。特别是低层 CLIP keys 特征,其对象结构信息损失严重,近乎无法识别。相反,所有层级的 DINO 特征都能跨越真实域与艺术域,准确清晰地分割并捕获对象的结构及其独特的语义成分(如人脸画像的头发、眼睛、鼻子等五官共享同样的纹理并区别于其他部分),且背景干净,保持了输入图像的结构特征。

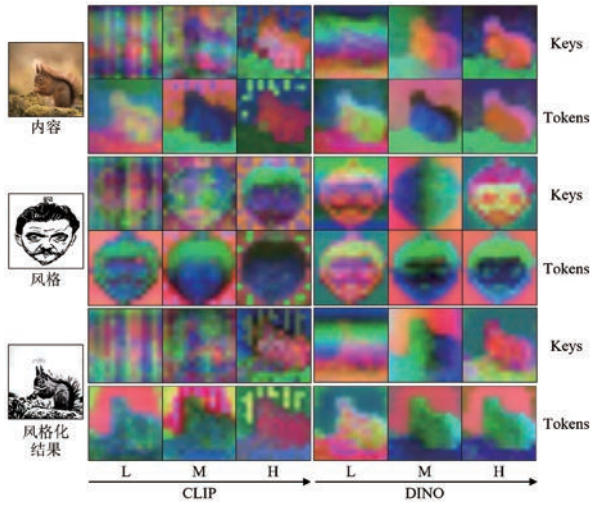


图 3　ViT 模型 CLIP[15]与 DINO[14]的分层特征可视化

本文认为,二者的差异主要源于它们各自不同的训练方式:DINO[14]是自监督训练模型,而 CLIP[15]是有监督训练模型。在理论层面上,DINO 认为传统的有监督训练(如 CLIP、原始 ViT[13]等)将图像中包含的丰富视觉信息压缩到只有类别的标签信息,不可避免地丢失图像中的细节信息以及没有被标签提及的目标信息。相反,DINO 的自监督训练能够根据图像的全局上下文信息创建任务并提供更加丰富的学习信号,从而能够充分关注图像的局部语义、形状与纹理等细粒度信息。此外,最近已有工作[12,34,56]表明,自监督 ViT 模型 DINO 的最后一个自注意力层的嵌入特征中包含显式的场景布局及物体边界信息,而这是 CLIP 和原始 ViT 等有监督 ViT 模型所缺乏的。

为了验证以上猜想,本文按照 Wysoczanska 等[56]的做法可视化对比了不同 ViT 模型的亲和图(Affinity Maps),它反映了特征 patch 之间的相关性。如图 4 所示,对于 DINO、CLIP 与原始 ViT 特征,我们逐 patch 计算特定种子特征(seed)与其他 patch 特征之间的余弦相似度并可视化为亲和图。按照文献[56],此处的 ViT 特征为最后一个自注意力层中丢弃了[CLS] tokens 后的 value 特征。从图 4 中可以看到,当计算特定 patch(黄色高亮点)与其余 patch 之间的相似度时,DINO 特征能够精准地匹配与查询点相似的语义区域(如第 1 组的鸟喙,第 2 组的钟楼红砖面,第 3 组的钟盘),并排除不相似的语义;而 CLIP 与原始 ViT 特征则带有大量噪声,对内容语义的区分度较低。由此可见,自监督 ViT 模型 DINO 具有良好的场景理解和语义定位能力,而 CLIP、原始 ViT 等有监督 ViT 模型则较为欠缺。

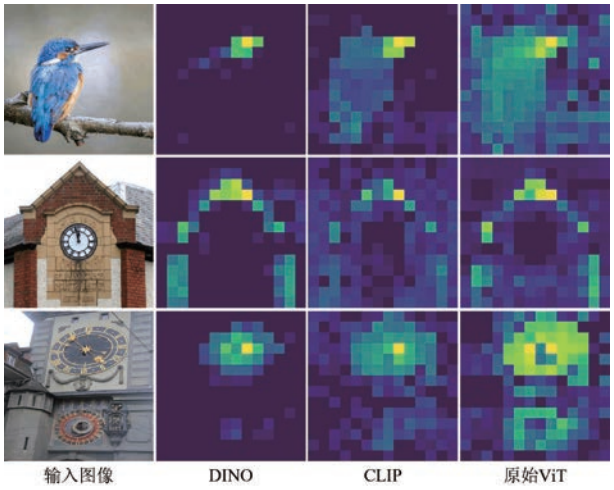


图 4　不同 ViT 模型的亲和图(Affinity Maps)对比(使用 DINO[14]、CLIP[15]与原始 ViT[13],计算种子特征(黄色高亮点)与其他 patch 特征之间的亲和度。DINO 特征能够精准地匹配与查询点相似的语义区域,并排除不相似的语义;而 CLIP 与原始 ViT 特征带有大量噪声,内容语义区分度较低。)

此外,针对艺术风格迁移任务而言,之前引入预训练 ViT 模型的方法[60-61]更多强调风格化效果,通过引入 CLIP 作为风格编码器或风格损失来加强艺术表现力,侧面证明了 CLIP 更擅长于学习风格图

像。而与现有方法相反，本文从空间语义感知的角度出发，目标是实现内容结构保持度高、相似语义区域风格化一致的艺术风格迁移。因此本文引入了具备空间感知特性的 DINO，并构造语义结构损失指导风格化结果的内容维持。最后，与 Zhou 等[34]一样，本文发现 keys 在结构语义保持方面优于 tokens，并通过消融实验对其进行了验证，故本文选用 DINO 的 keys 作为 ViT 指导的特征表示。

3.2.2 语义结构损失

基于 3.2.1 节对 ViT 语义指导的分析，本文构造了一种指导跨域艺术风格迁移的结构损失。此前，在图像翻译任务中，Tumanyan 等[12]利用 DINO 实现了语义对齐的自然图像之间的语义外观风格迁移，本文将其思路引入了非对齐的艺术风格迁移中。

对于内容的语义结构，艺术风格迁移需要一种对颜色、纹理、笔触等风格变化具有鲁棒性的表示，这种表示能够在跨域的迁移过程中保留前景对象及背景空间的边缘轮廓、场景布局和景深等语义信息。为了应对该挑战，本文将方法诉诸自相似性描述符——之前图像检索的经典方法[62]以及最近风格迁移中利用深度学习的方法[12,63]都证明了：自相似性描述符能够捕获域不变的结构信息（内容），而不受特定（风格）域的影响。因此，本文使用 DINO 的 keys 特征之间的自相似性作为结构表示。具体来说，首先从 DINO 中提取深层空间特征的 keys，然后使用余弦相似度计算它们之间的自相似性：

$$S^L(I)_{ij}=\frac{k_i^L(I)\cdot k_j^L(I)}{\|k_i^L(I)\|\ \|k_j^L(I)\|} \tag{1}$$

其中，I 是输入图像，L 为 ViT 提取特征的层，k 表示 keys，i 和 j 为不同的 keys 索引序号，S 是自相似性描述符（维度为 $S^L(I)\in\mathbb{R}^{(n+1)\times(n+1)}$，其中 n 是 ViT 的分块数量）。

如图 2 所示，在获得基于 DINO 特征的结构表示后，本文构造了一种语义结构损失 L_{vit}^{str} 以鼓励风格化图像 I_{CS} 与内容图像 I_C 的结构相匹配，该损失由 DINO 最深层（$L=12$）的注意力模块中提取并计算的 keys 自相似性差异来定义：

$$L_{vit}^{str}=\|S^L(I_{CS})-S^L(I_C)\|_F \tag{2}$$

其中，$S^L(I)$ 由式(1)定义，F 表示 F 范数。

3.2.3 知识蒸馏损失

为了直观地理解预训练 ViT 模型与传统 CNN 模型在特征提取方面的差异，本文通过 PCA[59]（主成分分析，主成分个数为 3）可视化了 DINO ViT[14]特征与 CNN[8]特征，如图 5 所示。可以看到，得益于 DINO 的空间感知特性，ViT 特征的对象边界清晰、结构层次分明，精准地分割了不同的内容语义；而 CNN 特征的对象边界模糊、前景与背景混杂（如第 3 列的人脸、第 4 列的猫），对内容语义的感知与分割存在误差（如第 1 列的鸟、第 2 列的大桥）。

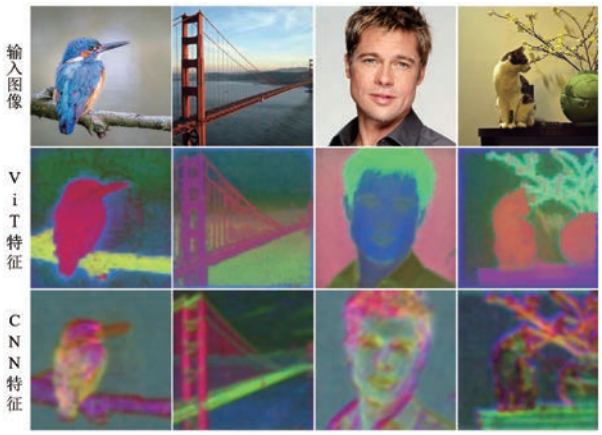


图 5 知识蒸馏路径中 ViT 特征与 CNN 特征的可视化

因此，为了从预训练的 DINO 中提炼出更通用的知识以加强编码器的特征提取能力，受到最近小样本生成任务中知识蒸馏发展的启发[38]，本文设计了一种知识蒸馏（Knowledge Distillation，KD）损失 L_{vit}^{kd}，如图 2 所示。L_{vit}^{kd} 会强制编码器的特征空间模仿 DINO 的特征空间，即让知识从图像的 DINO 特征提取到图像的编码器特征。具体来说，首先令内容图像 I_C 与风格化图像 I_{CS} 通过 DINO V 与编码器 E，分别提取出 DINO 特征 f_C^{vit}/f_{CS}^{vit} 和编码器特征 f_C^{cnn}/f_{CS}^{cnn}，然后通过 L1 损失约束二者的差异，迫使编码器的特征空间向 DINO 的靠近：

$$L_{vit}^{kd}=\|f_C^{cnn}-f_C^{vit}\|_1+\|f_{CS}^{cnn}-f_{CS}^{vit}\|_1 \tag{3}$$

注意，由于 DINO 和编码器的网络架构不同（DINO 是 ViT 架构，编码器是 CNN 架构），因此他们的特征形状并不一致：DINO 提取的是空间一维（1D）的 ViT 特征 $f^{vit}\in\mathbb{R}^{n\times d}$（$n$ 是除去[CLS] token 后的 ViT 分块 tokens 数量，d 是每个 tokens 的嵌入维度），而编码器提取的是空间二维（2D）的 CNN 特征 $f^{cnn}\in\mathbb{R}^{h\times w\times c}$（$h$ 与 w 代表特征图的高与宽，c 是特征图每个位置的通道数量）。

为了解决这个问题，本文针对 DINO 设计了一个 1D 转 2D 的 ViT 特征反转模块，以对齐 DINO 特征与编码器特征的空间尺寸。具体来说，本文反转了将图像输入 ViT 时的分块操作：首先，将一维 ViT 特征按输入时的分块顺序重排为二维特征；然后，使用一个卷积核大小为 ViT 分块尺寸的转置卷

积将重排特征的空间尺寸缩放至与编码器特征的空间尺寸一致;接着,通过一个 1×1 卷积转换重排特征的通道数量,使其在通道维度上与编码器特征一致。最后输出二维的反转 ViT 特征 $f^{vit}\in\mathbb{R}^{h\times w\times c}$。在匹配了 DINO 特征与编码器特征的维度后,本文通过式(3)计算基于 DINO 的知识蒸馏损失。

3.3　空间结构损失和频域纹理损失

3.3.1　空间结构损失

现有方法通常会采用基于 VGG 特征的感知损失来约束编码器/解码器。然而,如果没有显式的结构先验指导,风格化图像可能难以捕获内容的结构、布局及景深等层次关系[50]。为此,本文增强了编码器/解码器对空间结构信息的感知能力。具体来说,如图 2 所示,首先利用拉普拉斯算子[16]将输入图像滤波为边缘表示。该算子滤除了色彩、笔触、纹理等特定区域的冗余风格信息,只保留了物体轮廓、场景布局等各种层次的空间结构信息,天然有利于在风格迁移任务中保持内容语义。同时,与传统的像素级损失相比,关注高频成分的拉普拉斯算子能够有效避免对背景区域的过多干扰,从而更加突出对关键内容和场景布局的结构保持。此外,它对小尺度变化具有良好的敏感性,可以使结果图像在局部细节上更加逼真,有助于解决本任务的内容跨域一致性保持问题。拉普拉斯算子由具有拉普拉斯核的单个卷积层组成:

$$Kernel_{Laplace}=\begin{bmatrix}0&-1&0\\-1&4&-1\\0&-1&0\end{bmatrix}\tag{4}$$

在获得边缘表示后,构造空间结构损失 L_{lap}^{str} 与空间重构损失 L_{lap}^{rec}。前者在风格化路径中约束风格化图像 I_{CS} 与内容图像 I_C 之间全局结构的差异,后者在重构路径中约束重构图像 $\widetilde{I}_C$ 与内容图像 I_C 之间的差异:

$$L_{lap}^{str}=\|Lap(I_{CS})-Lap(I_C)\|_1\tag{5}$$

$$L_{lap}^{rec}=\|Lap(\widetilde{I}_C)-Lap(I_C)\|_1\tag{6}$$

$$L_{lap}=L_{lap}^{str}+L_{lap}^{rec}\tag{7}$$

其中,Lap 表示拉普拉斯运算[16]。

3.3.2　频域纹理损失

为了减小模型的频率偏差[17,20],使其在图像重构过程中重视高频信息,本文进一步对提取的特征进行 Haar 小波变换[22],获得输入图像的高频信号。Haar 小波能够通过小波池化将图像转换到小波域,包含四个卷积核:LL^T,LH^T,HL^T,HH^T,其中 $L^T=\frac{1}{\sqrt{2}}[1,1]$,$H^T=\frac{1}{\sqrt{2}}[-1,1]$,$L$ 和 H 分别表示低通滤波器和高通滤波器。低通滤波器 L 捕获图像的轮廓和表面,而高通滤波器 H 侧重于边缘和纹理等细节信息[20]。此外,与传统的傅里叶变换不同,小波变换具有良好的局部化性质,能够在空间域中准确捕捉图像的细节纹理信息。这一特性对于实现区域一致性的风格化非常重要,因为其不仅要求保留图像的整体结构,还要求保留局部语义的细节信息。而小波变换可以在不丢失空间信息的前提下提取图像的频率特征,有助于准确地重建内容结构。

图 6 展示了通过 Haar 小波获得的给定图像的频率分量。如图 6 所示,低频分量 LL 捕获了图像的整体语义,而高频分量(即 LH、HL、HH)包含更多细粒度信息。此外,通过将三个高频分量相加,能够大致获得图像的所有细节信息(如女人的眼睛和嘴唇)。基于这种频域纹理表示,如图 2 所示,本文构造了频域纹理损失 L_{fre}^{tex} 以及频域重构损失 L_{fre}^{rec},以减少风格化结果的内容细节丢失:

$$L_{fre}^{tex}=\|Haar(I_{CS})-Haar(I_C)\|_1\tag{8}$$

$$L_{fre}^{rec}=\|Haar(\widetilde{I}_C)-Haar(I_C)\|_1\tag{9}$$

$$L_{fre}=L_{fre}^{tex}+L_{fre}^{rec}\tag{10}$$

其中,$Haar$ 表示将特征分解为不同频率分量的 Haar 小波变换[22]。

图 6　Haar 小波变换[22]后的频率分量示意图

3.4　训练优化

如图 2 所示,本文以端到端的方式训练编码器和解码器执行风格化任务及重构任务。

重构任务将风格化图像 I_{CS} 作为内容、内容图像 I_C 作为风格,强制模型重构原始内容图像 $\widetilde{I}_C$。除了 3.3 节中提出的空间重构损失 L_{lap}^{rec} 和频域重构

损失 L_{fre}^{rec}，本文还遵循基线方法 CAP-VST[9]，采用了 L1 范数计算的循环一致性像素重构损失[54]：

$$L_{pix}^{rec}=\| \widetilde{I}_C - I_C \|_1 \tag{11}$$

在风格化任务中，遵循基线方法[9]，本文采用了抠图拉普拉斯损失(Matting Laplacian loss)[9]，这是一种源于经典图像抠图任务的正则化方法[64]，用于平滑结果图像，抑制失真现象。其定义如下：

$$L_{mat}=\frac{1}{N}\sum_{c=1}^{3}V_c[I_{CS}]^T M V_c[I_{CS}] \tag{12}$$

其中，N 表示图像像素的数量，$V_c[I_{CS}]$ 表示风格化图像 I_{CS} 在通道 c 上的矢量化，M 表示内容图像 I_C 的抠图拉普拉斯矩阵。

对于风格约束，本文采用了经典风格损失[3]以使结果图像的风格向参考图像的风格靠近：

$$\begin{aligned}L_{style}=&\sum_{i=1}^{l}\| \mu(\phi_i(I_{CS}))-\mu(\phi_i(I_S)) \|_2\\&+\sum_{i=1}^{l}\| \sigma(\phi_i(I_{CS}))-\sigma(\phi_i(I_S)) \|_2\end{aligned} \tag{13}$$

其中，I_S 表示风格图像，ϕ_i 表示 VGG-19 网络[8]的第 i 层(从 $ReLu1\_1$ 到 $ReLu4\_1$)，μ 和 σ 分别表示特征图的均值和方差。

综上所述，本文方法的总损失函数定义如下：

$$\begin{aligned}L_{total}=&\lambda_{vit}^{str}L_{vit}^{str}+\lambda_{vit}^{kd}L_{vit}^{kd}+\lambda_{lap}L_{lap}+\lambda_{fre}L_{fre}\\&+L_{mat}+L_{pix}^{rec}+L_{style}\end{aligned} \tag{14}$$

其中，λ 是对应于各损失的权重。

3.5 视频风格迁移

在视频风格迁移任务中，现有的单帧方法[9,29-30]已经表明，通过对每个视频帧单独应用图像风格迁移算法可以完成视频风格迁移。本文方法作为一种训练损失，可以加强单帧方法[9,29-30]对视频内在一致性的保持，维持内容视频自身的连贯性和稳定性，使风格化后的视频保持视觉稳定，避免帧间闪烁伪影。同时，得益于本文方法作为训练损失的即插即用特性，可以将其与复合时序正则化项[65](一种用于视频风格迁移的损失函数)无缝联合，在图像风格迁移的训练结束后共同微调骨干网络。

4 实验结果与分析

4.1 概述

本节围绕实验展开，共包含 8 个小节，内容安排如下：(1)实验细节。该节从 4 个方面阐述实验部分的细节设置，包含数据集、参数设置、对比方法、评价指标。(2)对比实验。该节为本文方法与现有方法的对比实验及分析，包含定性比较、定量比较、模型复杂度与效率分析三部分，旨在验证本文方法的有效性与先进性。其中，定性比较与定量比较均分为图像方法和视频方法两种对比实验。(3)消融实验。该节为本文方法关键设计与超参数设置的消融实验，旨在验证本文方法所涉及各模块的有效性与合理性，包含 8 个部分：空间结构损失与频域纹理损失、ViT 语义结构损失、ViT 知识蒸馏损失、ViT 特征表示的选择、预训练 ViT 模型的选择、边缘滤波算子的选择、边缘滤波算子的卷积核大小、小波变换的分解尺度。(4)本文方法在不同骨干网络上的应用。该节为本文方法在不同现有方法骨干网络上的应用，旨在验证本文方法的灵活性与兼容性。(5)与现有方法损失函数的对比实验。该节为本文所提出损失函数与现有损失函数的对比实验，旨在验证本文方法的有效性与优越性。(6)本文方法在图像翻译任务上的应用。该节为本文方法在其他图像生成任务(图像翻译任务)上的应用，旨在验证本文方法的跨任务适用性。(7)局限性。该节分析了本文方法现存的局限性，并展望了未来工作方向。(8)实验总结。该节归纳并总结了本文的实验结果及分析。

本节具体实验内容及结果分析如下列小节所示。

4.2 实验细节

(1)数据集。本文采用风格迁移任务的通用数据集——内容图像数据集 MS-COCO[23]与风格图像数据集 WikiArt[24]。MS-COCO(Microsoft Common Objects in Context)[23]是一个大规模(包含 164 K 张图像)的目标检测、分割和字幕数据集。该数据集以场景理解为目标，主要包含自然环境中各类常见物体的复杂日常场景，现已被广泛用作风格迁移任务的内容数据集。WikiArt 数据集[24]包含来自 1119 位艺术家创作的 81446 幅艺术作品，涵盖了从 15 世纪创作的艺术作品到 21 世纪创作的现代美术绘画。其中包含 27 种风格(抽象派、巴洛克、立体主义、印象派等)和 45 种体裁(城市景观、风景、肖像、静物等)，美学构成多样化。目前，WikiArt 作为风格数据集已被广泛应用于风格迁移任务。本文从 MS-COCO[23]和 WikiArt[24]中各自随机抽取了 80000 张图像作为内容和风格数据集。

在测试图像风格迁移时，本文在 MS-COCO[23]和 WikiArt[24]数据集中随机选择了 49 张内容图像和 48 张风格图像，总共生成了 2352 张测试结果。

在测试视频风格迁移时,本文采用 MPI Sintel[66]数据集:一个用于光流评估的开源数据集,现也常用于视频风格迁移的评估[9,27,29]。本文在 MPI Sintel[66]中随机选择了 10 段视频片段(每段 50 帧,12 FPS),并使用 10 张风格图像分别迁移这些内容视频。

(2)参数设置。本文采用 CAP-VST[9]作为基线风格迁移模型,并使用 Adam 优化器[67]对网络进行 160000 次迭代训练,batch size 为 2。初始学习率设置为 1e-4,并在 5e-5 处衰减。对于视频风格迁移,联合复合时序正则化项[65]再进行 10000 次迭代以微调模型。训练时,对输入的内容图像与风格图像进行预处理:将所有输入图像的较短边都重新缩放为 512 像素(保留长宽比),然后随机裁剪为 256×256 像素。测试时,由于本文网络是全卷积的,因此它可以处理任意输入大小的图像。所有实验均在 NVIDIA Tesla V100(32GB) GPU 上进行。本文将损失函数的权重参数设置为:$\lambda_{vit}^{str}=5$,$\lambda_{vit}^{kd}=1$,$\lambda_{lap}=1$,$\lambda_{fre}=1$。

(3)对比方法。对于图像风格迁移,本文与 13 种前沿艺术风格迁移方法进行比较,对所提出方法的有效性进行了定性和定量评估。对比方法包括:S2WAT[68]、AesFA[69]、CAP-VST[9]、TSSAT[28]、MicroAST[31]、StyTr2[70]、EFDM[25]、CCPL[30]、StyleFormer[71]、IECST[29]、AdaAttN[27]、ArtFlow[10](包括 AdaIN 与 WCT 两种版本,记作 ArtFlow-A 与 ArtFlow-W)、CKD[7]。其中,EFDM[25]、TSSAT[28]、AdaAttN[27]使用 VGG-19[8]作为骨干网络,并设计了基于全局统计分布或局部注意力机制的特征转换方法。IECST[29]和 CCPL[30]在此基础上引入了对比学习策略[32]以进一步加强生成质量。为了实现超分辨率的风格迁移,CKD[7]针对 VGG-19 提出了一种知识蒸馏方法,MicroAST[31]设计了一种微编码器-解码器网络,AesFA[69]在前者的基础上引入了八度卷积[72]。为了缓解内容泄漏问题,CAP-VST[9]与 ArtFlow[10]利用可逆神经流模型[11]替换了 VGG-19 作为骨干网络。除了以上基于 CNN 的方法,本文还对比了基于 Transformer 的方法:StyleFormer[71]首先在 VGG-19 的范式上引入了 Transformer 驱动的参数化风格组合机制;StyTr2[70]则直接基于 Vision Transformer (ViT)[13]架构设计了风格迁移网络以实现无偏的内容表示;S2WAT[68]进一步通过分层 ViT 架构提出了条状窗口注意力 Transformer,有效地缓解了局部性问题。

对于视频风格迁移,本文与 6 种前沿方法进行比较,包括:UniST[73]、CAP-VST[9]、IECST[29]、AdaAttN[27]、MCCNet[74]、ReReVST[65]。其中,UniST[73]、MCCNet[74]、ReReVST[65]是专门为视频风格迁移任务设计的算法:UniST[73]设计了域交互 Transformer 框架,MCCNet[74]提出了多通道相关性网络,而 ReReVST[65]引入了复合时序正则化方法。

对比方法的结果均是通过重新训练各个方法开源发布的代码获得的(使用默认配置)。

(4)评价指标。

(i)图像风格迁移。为了全面评估不同算法的性能,本文按照之前的方法[9,69-70]采用了多种指标来评估结果的内容保持程度、风格化效果与综合视觉质量。

对于内容保持,本文采用 SSIM[75]、LPIPS[76]以及 Content Loss[3]对结果进行评估,具体如下:

① SSIM (Structural Similarity Index)[75]:SSIM 通过模拟人眼对图像结构信息的感知方式,从像素层面综合评估结果图像与内容图像在亮度、对比度和结构三方面的一致性,分数越高越好(↑)。

② LPIPS (Learned Perceptual Image Patch Similarity)[76]:LPIPS 是一种感知相似性评价指标,用于衡量两幅图像在感知层面上的差异。与 SSIM 等传统指标不同,LPIPS 通过深度神经网络来模拟人类视觉系统对图像相似度的感知,能够捕捉高层语义变化(如纹理、结构改变),更加贴近人类主观感受。本节 LPIPS 指标通过预训练神经网络 AlexNet[77]的中间特征表示衡量风格化图像与内容图像之间的感知相似性,分数越低越好(↓)。

③ Content Loss[3]:Content Loss 是风格迁移任务中常用的损失函数,基于预训练的深度神经网络 VGG-19 衡量风格化图像与原始内容图像在高层语义特征空间中的相似程度,反映了包含物体结构、轮廓和布局在内的图像内容是否保持一致,分数越低越好(↓)。

对于风格化效果,本文采用 Style Loss[3]对结果进行评估,具体如下:

① Style Loss[3]:Style Loss 是一种衡量风格化图像与目标风格图像的风格相似程度的指标。它关注图像的纹理、笔触、颜色等视觉风格特征,通过预训练深度神经网络(VGG-19)提取图像的不同中间层特征并计算 Gram 矩阵以衡量风格的一致性。其中,Gram 矩阵能够反映特征图不同通道之间的相

关性,捕捉图像的纹理、颜色分布等风格信息。Style Loss 的分数越低越好(↓)。

对于综合视觉质量,本文采用 CLIP-IQA[78] 和 SSIM-CLIP 指标对结果进行评估,具体如下:

①CLIP-IQA[78]:CLIP-IQA 是一种基于 CLIP[15] 的 BIQA(Blind Image Quality Assessment,无参考图像质量评估)方法,旨在模拟人类对图像"观感"的主观判断。它利用 CLIP 中封装的丰富视觉-语言知识与强大的图文对齐能力,通过与预设文本提示的相似度计算,直接评估风格化图像的视觉感知质量,分数越高越好(↑)。

②SSIM-CLIP:考虑到风格化质量取决于内容保持度和风格匹配度的平衡表现,本文受到 ArtFID[79] 思想的启发,设计了 SSIM-CLIP 指标以综合评估风格化图像。具体来说,SSIM-CLIP 采用 SSIM[75] 作为内容评估因子,CLIP Loss[15] 作为风格评估因子,结合二者进行计算。本文观察到最近基于扩散模型的方法[60-61] 大多使用 CLIP[15] 作为风格编码器或风格损失来指导风格化过程,因此本节采取了 CLIP Loss[15] 作为风格评估因子。SSIM-CLIP 的计算公式如下,结果分数值越小越好(↓):

$$SSIM-CLIP(I_{cs},I_c,I_s)=(2-SSIM(I_{cs},I_c))\cdot(1+CLIPLoss(I_{cs},I_s)) \tag{15}$$

其中,I_{cs},I_c,I_s 分别表示结果图像、内容图像与风格图像。为了平衡内容因子与风格因子的重要程度,同时避免 $I_{cs}=I_c$ 或 $I_{cs}=I_s$ 时计算分数退化到最小值 0,本文将常数 2 和 1 分别添加到内容因子和风格因子中以缩放二者尺度,确保只有同时在内容保持和风格匹配方面表现良好的方法才能获得较优的 SSIM-CLIP 分数。

(ii)视频风格迁移。本文参考之前的方法[9,65],采用 ReReVST[65] 中定义的时间损失(Temporal loss)来衡量短期(帧间隔 $i=1$)和长期(帧间隔 $i=10$)的时间一致性。此外,与图像方法的评估指标不同,针对视频方法,本文使用 LPIPS 计算风格化视频帧之间短期($i=1$)和长期($i=10$)的变化程度,以反映视频结果的一致性和稳定性。同时,本文逐帧计算风格化视频与内容视频/风格图像之间的内容损失(Content Loss)和风格损失(Style Loss),作为评估内容保持与风格化效果的指标。

4.3 对比实验

4.3.1 定性比较

(1)图像风格迁移。图 7 展示了定性比较的直观结果。基于流模型[11] 的方法特征表示能力有限,因此 ArtFlow[10] 的结果普遍存在风格化不一致(如第 3 行的天空区域)或色彩偏差(如第 4 行的整体色调)等问题。AdaAttN[27] 将 SANet[26] 与 AdaIN[3] 结合,成功改善了前者的风格-注意力机制在内容结构保持方面的不足,但也导致了风格退化问题,使得生成结果的色彩黯淡(如第 6~8 行的画面色彩),且依旧存在背景纹理杂乱等局部语义区域的风格化不一致现象(如第 6 行的天空区域)。得益于 GAN[80] 良好的图像生成能力,IECST[29] 的视觉质量优于其他方法。但是由于 GAN 的外部训练策略,使得其生成结果的风格会偏离输入图像的内部参考风格(如第 6~8 行的背景区域)。EFDM[25] 与 TSSAT[28] 分别通过精准特征分布匹配和两阶段特征分布转换提升了生成结果的风格表现力,然而两种方法都忽略了对内容结构跨域一致性的保持,导致生成结果的内容结构受到严重破坏(如第 1 行的熊和第 2 行的松鼠)。

AesFA[69] 和 CAP-VST[9] 分别引入了八度卷积[72] 和可逆神经网络[11] 以替换 VGG-19 编码器/解码器,但它们的特征提取能力仍然有限,会产生内容细节丢失(如第 2 行的松鼠)以及风格化不一致(如第 3 行的天空)的结果。CCPL[30] 引入了对比学习策略[32] 以保持来自内容源的细节信息,但为此牺牲了艺术风格化效果,丢失了笔触、纹理等体现艺术作品质感的元素(如第 5~8 行的结果)。此外,尽管 CCPL 设计的简易协方差变换(SCT)模块提高了推理效率,但过于简单的风格转换可能导致生成结果在物体边缘出现红蓝条带状伪影(如第 5 行的人脸风格化结果)。CKD[7] 与 MicroAST[31] 分别通过知识蒸馏[35] 与微编码器-解码器构建了轻量级网络以替换 VGG-19,实现了超分辨率风格迁移。但是二者受限于轻量级网络的设计,均无法有效地表征内容或风格信息:CKD[7] 无法保持内容结构,图像的内容信息丢失严重(如第 1~3 行的结果);相反,MicroAST[31] 不能有效编码风格化信息,生成结果的风格化程度较低。

在基于 Transformer 的方法当中,StyleFormer[71] 将基于 CNN 的编码器-解码器架构与 Transformer 驱动的风格组合模块相结合,试图实现细粒度风格多样性和语义内容连贯性。但其受限于由有限风格代码构成的风格库策略,生成结果缺乏风格多样性,且无法捕获细粒度的参考风格元素(如第 6~7 行的整体色调与背景的风格纹理)。StyTr2[70]

图 7　与不同图像风格迁移方法的定性比较(各个方法的发表会议与年份通过斜体标注)

首次基于完整的 Vision Transformer (ViT)[13]架构设计了风格迁移网络,利用 Transformer 的长距离依赖特性有效提取和保持了输入图像的全局内容信息。S2WAT[68]基于分层 ViT 架构提出了一种条状窗口注意力 Transformer 以同时捕获局部和全局信息,进一步保持了内容细节。但以上两种方法都因此而过度削弱了结果的风格化表现力,使其整体色彩和局部纹理都与参考图像有所差距(如第 5 行与第 8 行的结果)。

相反,本文方法通过 DINO 语义指导以及空间结构和频域纹理的感知增强,保持了内容语义的跨域一致性,有效缓解了内容细节丢失与局部语义区域风格化不一致的问题(如第 1～4 行的结果),同时没有过度牺牲艺术表现力(如第 5～8 行的结果),实现了内容保持与风格化的平衡。

此外,由于本文采用全卷积网络,因此所提方法能够处理任意分辨率,包括超高清分辨率(如 4K)的输入图像。如图 8 所示,即使在处理超高分辨率的内容图像时,本文方法也能够在迁移风格元素(如色调、笔触及纹理)的同时保持内容结构不失真及一致性风格化。相比之下,现有方法很难同时兼顾这两方面。例如,CAP-VST[9]能够生成像艺术作品一样的结果,但仍然面临局部语义区域风格化不一致的问题(图 8 中的天空区域)。另一方面,尽管 AesFA[69]、MicroAST[31]、CCPL[30]和 CKD[7]通过八度卷积[72]、微编码器-解码器[31]、简易协方差变换[30]或知识蒸馏[35]等方法轻量化了网络,实现了实时超高清分辨率风格迁移,但这些方法受限于自身过于简化的模型结构,网络缺乏对深层特征的表征能力,生成结果均牺牲了风格化效果,失去了真实艺术作

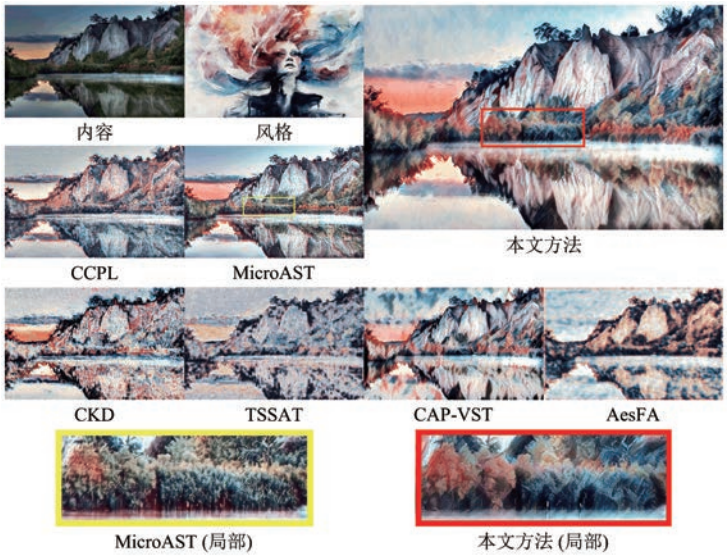


图 8 超高清分辨率(4K：4096)的结果对比

品的质感(图 8 中黄色框)。相反,本文方法利用了 DINO-ViT[14]蕴藏的大模型知识与可逆神经网络[11]的信息保持能力,实现了兼具内容保持与风格化表现力的输出图像(图 8 中红色框)。

同时,为了进一步证明本文方法的优越性,按照文献[50,81]的做法,我们可视化了不同方法风格化结果之间的边缘图与深度图,如图 9 所示。理想情况下,艺术风格迁移应该在充分保留内容结构、层次和景深的基础上呈现参考风格的色调、笔触和纹理。这意味着风格化结果的边缘图和深度图应该与原始

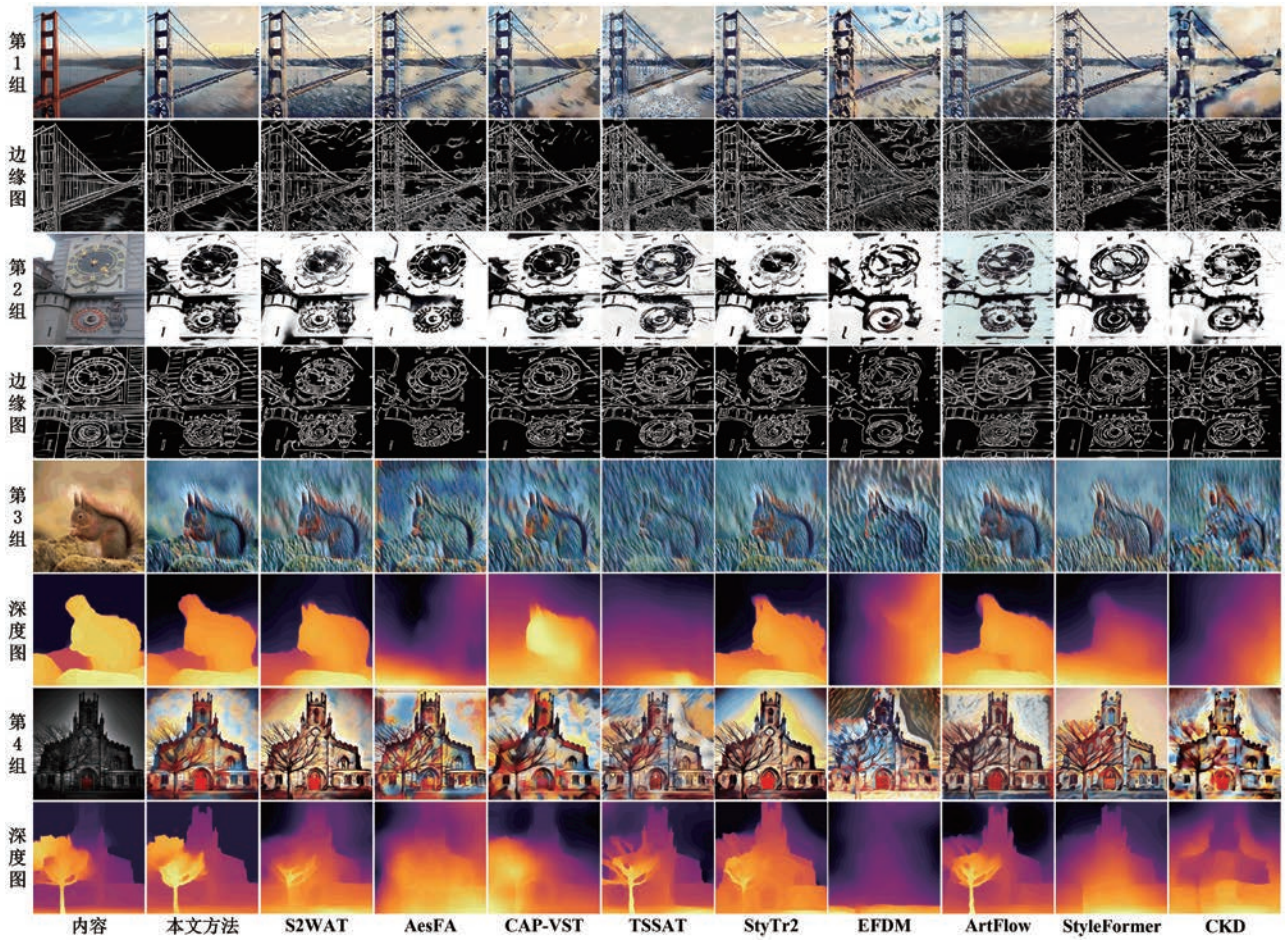


图 9 不同方法风格化结果的边缘图及深度图可视化对比

内容的保持一致。为了进行准确的比较,我们使用最近的 LDC[82] 和 MiDaS[83] 方法分别生成结果的边缘图和深度图。如图 9 所示,在边缘图实验中,现有方法会在风格化结果的背景区域引入不必要的纹理伪影,并且不同程度地破坏前景物体的边缘轮廓(如第 1 组)。哪怕是在内容保持方面具有显著优势的 Transformer 类方法[68,70-71] 和流模型方法[9-10],都不可避免丢失参考图像的结构信息(如第 2 组)。相比之下,本文方法的边缘图与原始内容的最为接近,保持了局部语义区域的风格化一致性和内容层次的完整性。在深度图实验中,AesFA[69]、TSSAT[28]、EFDM[25] 及 CKD[7] 的结果都严重破坏了内容源的景深信息,已经无法完成深度估计(如第 3 组)。同时,通过第 4 组实验可以看出,与 S2WAT[68]、StyTr2[70] 和 StyleFormer[71] 等 Transformer 类方法相比,本文方法不仅有效保持了内容图像的层次与景深,还具备更加富有表现力的风格化效果。

(2)视频风格迁移。为了直观地比较视频结果的稳定性和一致性,我们在图 10 中可视化了连续帧之间的时间误差热图[9],它反映了模型在短期(相邻两帧之间)的时间一致性保持能力。可视化误差热图时,我们使用 MPI Sintel 数据集[66] 提供的真实光流图作为基准。如图 10 所示,ReReVST[65] 采用的复合正则化过度强调对时序稳定性的保持,牺牲了风格化效果导致整体色彩退化。MCCNet[74] 则忽略了对视频内容结构的维护,会产生对象轮廓扭曲失真的结果。受到 GAN[80] 训练的限制,IECST[29] 在视频风格迁移时依然会产生与参考图像的色调偏差较大的结果,同时其帧间时间误差也较大。AdaAttN[27] 输出了迁移良好的结果,但在色彩表现和细节维护上仍有改进的空间。UniST[73] 基于 Transformer 设计的联合学习框架成功保持了视频结果的时序稳定性,但与 ReReVST[65] 一样,该方法严重限制了风格化表达,同时模糊了视频的内容结构,生成结果的视觉质量较差。采用了可逆神经网络[11] 的 CAP-VST[9] 在各方面都取得了较为均衡的结果,而本文方法在其基础上进一步增强了运动帧稳定性(如蓝框所示,时间误差热图与原始视频的更接近),提升了细节内容语义保持度(如红框所示,背景墙面纹理与原始视频的更接近),同时避免了艺术效果过度损失。

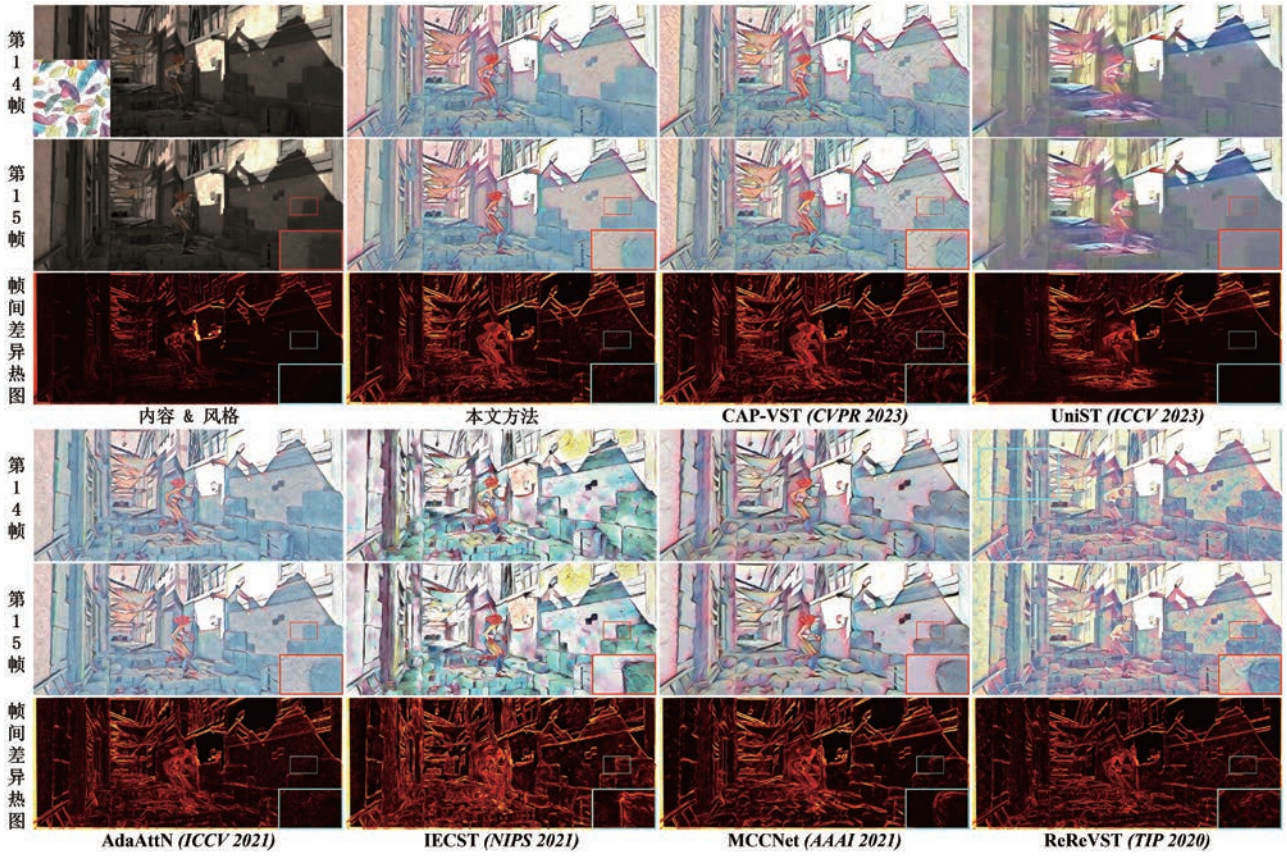


图 10　与不同视频风格迁移方法的定性比较(各个方法的发表期刊(会议)与年份通过斜体标注。每一组中,前两行为相邻帧,最后一行为相邻帧之间的差异热图,即时间误差热图[9]。)

我们认为这可以归因于三点：(1)本文的“DINO-ViT+自相似性嵌入”策略使模型在训练时更倾向于关注图像之间的语义结构相似性，从而避免对风格化表达产生负面影响；(2)边缘滤波、小波变换与重构损失的结合提供了适当的引导，在细粒度层面提升了视频结果的内容保真度和时序稳定性；(3)知识蒸馏的引入让网络能够联合预训练 ViT 大模型与可训练神经流模型的能力，增强了结果的视觉质量。

4.3.2 定量比较

表 1 展示了与不同 SOTA 图像风格迁移方法的定量比较，最优的用粗体表示，次优的用下划线表示，本文方法相较基线模型 CAP-VST[9] 取得的指标提升(↑)或下降(↓)用 Δ 基线表示。如表 1 所示，许多内容保持优秀的方法(如 MicroAST[31]、IECST[29] 及 AdaAttN[27] 等)都过度牺牲了风格化效果(很高的 Style Loss 值)。另一方面，一些强调风格化的方法(如 AesFA[69]、CKD[39] 等)则忽略了对内容结构的保持(较低的内容指标分数)。得益于可逆神经网络和 Transformer 的无偏全局建模特性，CAP-VST[9]、S2WAT[68] 及 StyTr2[70] 等方法整体表现良好。本文结合二者优势，通过 ViT 语义结构和知识蒸馏双指导的方式联合预训练 Transformer 大模型与可逆神经网络，不仅在内容保持方面取得了显著优势(SSIM 最优，LPIPS、Content Loss 次优)，同时将风格化效果维持在了较优区间(Style Loss 值与 AesFA[69]、StyTr2[70]、CKD[39] 的接近)，最终达到了整体最优的视觉质量(CLIP-IQA、SSIM-CLIP 最优)。此外，从增量的角度而言(如 Δ 基线所示)，在应用了本文所提各项损失函数后，基线模型 CAP-VST[9] 在所有内容保持指标(SSIM、LPIPS、Content Loss)与整体视觉质量指标(CLIP-IQA、SSIM-CLIP)上均取得了显著提升，而风格化指标仅有小幅下降(Style Loss)，证明了本文方法的有效性与应用性。在更多现有方法上的应用实例将在 4.5 节展示。

表 1 与不同图像风格迁移方法的定量比较

方法	发表会议/期刊及年份	SSIM↑	LPIPS↓	Content Loss↓	Style Loss↓	CLIP-IQA↑	SSIM-CLIP↓
本文方法	—	**0.607**	0.600	3.348	0.628	**0.386**	**1.931**
Δ 基线[9]	CVPR 2023	+0.079	−0.096	−1.035	+0.105	+0.024	−0.056
S2WAT[68]	AAAI 2024	0.398	0.619	3.789	0.537	0.347	2.215
AesFA[69]	AAAI 2024	0.407	0.686	4.209	0.587	0.341	2.123
CAP-VST[9]	CVPR 2023	0.528	0.696	4.383	**0.523**	0.362	1.987
TSSAT[28]	ACM MM 2023	0.361	0.665	3.428	0.846	0.373	2.212
MicroAST[31]	AAAI 2023	0.448	0.619	3.508	0.832	0.339	2.155
StyTr2[70]	CVPR 2022	0.431	0.619	3.777	0.592	0.362	2.164
CCPL[30]	ECCV 2022	0.436	0.621	**3.331**	0.561	0.335	2.192
EFDM[25]	CVPR 2022	0.312	0.710	4.792	0.879	0.369	2.281
StyleFormer[71]	ICCV 2021	0.398	0.645	3.781	0.705	0.384	2.212
ArtFlow-W[10]	CVPR 2021	0.439	0.614	3.735	0.737	0.363	2.132
ArtFlow-A[10]	CVPR 2021	0.405	0.633	3.723	0.732	0.367	2.170
AdaAttN[27]	ICCV 2021	0.493	0.610	3.652	0.929	0.374	2.075
IECST[29]	NIPS 2021	0.373	**0.581**	3.601	0.849	0.373	2.168
CKD[7]	CVPR 2020	0.269	0.748	5.512	0.597	0.316	2.284

注：粗体表示最优值，下划线表示次优值. ↑/↓：表示指标的衡量方向，“↑”表示该项指标的分数越高越好，“↓”表示该项指标的分数越低越好。+/−：表示指标分数的相对增加(+)或减少(−). Δ 基线[9]：将本文方法应用至基线模型 CAP-VST[9] 后，指标分数的相对增加(+)或减少(−)。

为了进一步直观地对比不同方法的风格化质量，本文按照 Ke 等[84] 的做法可视化了定量指标：采用 LDC 模型提取结果图像与内容图像的边缘图并计算 SSIM 分数作为内容相似度，同时利用 Ke 等[84] 提出的专为风格识别而训练的判别器来预测结果图像与风格图像的风格相似度。定量指标可视化如图 11 所示。直观上，内容相似度和风格相似度越高，风格化的整体视觉质量就越高。理想情况下的风格迁移方法应趋近于“理想点”，因此我们以本文方法到“理想点”的距离为半径绘制了等效质量曲线，该曲线上的所有点可以视为具有等效风格化质量。可以看到，本文方法的内容相似度最高，且风格相似度仅次于 CAP-VST[9] 与 CKD[39]，最终在整体风格化质量上最接近“理想点”，实现了内容保留和风格化效果之间的优效平衡。

视频风格迁移：如表 2 所示，本文方法在时间一致性、内容保持度及风格化效果三个方面均取得了

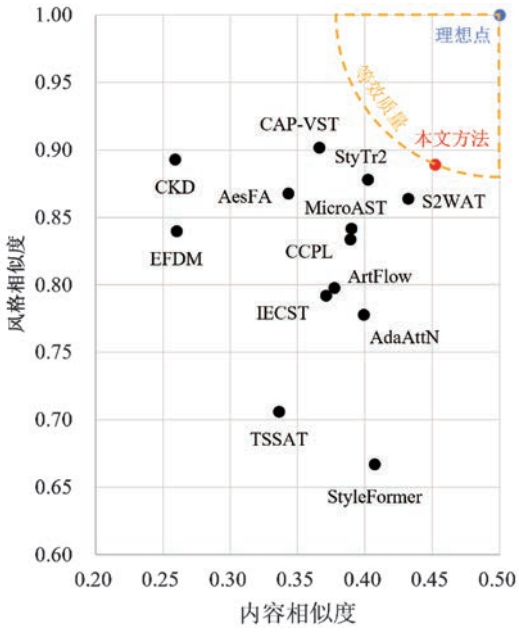


图 11　与不同图像风格迁移方法的定量比较(直观上,内容相似度和风格相似度越高,风格化的整体视觉质量就越高。理想情况下风格迁移方法应趋近于"理想点"。以本文方法到"理想点"的距离为半径绘制等效质量曲线,该曲线上的所有点可以视为具有等效风格化质量。)

最优或次优的指标评分,表明了本文方法的有效性与平衡性。值得注意的是,如 ΔUniST 所示,相较最近基于 Transformer 的 SOTA 视频风格迁移方法 UniST[73],本文方法在保持时间一致性评分相当的同时,具有远超它的风格化分数(与图 10 定性分析的趋势相同),证明了本文方法的优越性。

4.3.3　模型复杂度与效率分析

如表 3 所示,本文进行了不同方法的模型复杂度分析。为了全面分析,我们同时测试了训练阶段与推理阶段的模型复杂度。所有方法均在单张 NVIDIA Tesla V100(32GB) GPU 上测试,推理测试时图像的分辨率为 512×512。从表 3 中可以看到,由于本文设计的是训练阶段的损失函数(仅在训练时调用 ViT 大模型),所以不会增加基线模型[9]在推理阶段的模型复杂度与推理时间,证明了本文方法可以作为训练插件兼容于任意实时风格迁移模型的有效性与灵活性。同时可以观察到,现有的 Transformer 类方法(如 S2WAT[68]、StyTr2[70])普遍具有很高的模型复杂度和很低的推理效率,它们为了风格化效果牺牲了一定的实时性能,这是 Transformer 本身的高计算成本特性所致。

表 2　与不同视频风格迁移方法的定量比较

方法	发表会议/期刊及年份	Temporal Loss↓		LPIPS↓		Content Loss↓	Style Loss↓
		$i=1$	$i=10$	$i=1$	$i=10$		
本文方法	—	**0.164**	0.258	0.193	0.447	**2.576**	0.656
Δ 基线[9]	CVPR 2023	−0.006	−0.005	−0.006	−0.004	−0.335	+0.052
ΔUniST[73]	ICCV 2023	−0.001	−0.001	+0.005	−0.003	−0.627	−0.683
UniST[73]	ICCV 2023	0.165	**0.257**	**0.188**	0.450	3.203	1.339
CAP-VST[9]	CVPR 2023	0.170	0.263	0.199	0.451	2.911	**0.604**
AdaAttN[27]	ICCV 2021	0.187	0.271	0.230	0.479	3.689	0.779
IECST[29]	NIPS 2021	0.225	0.318	0.277	0.492	4.279	0.700
MCCNet[74]	AAAI 2021	0.176	0.264	0.218	0.450	3.879	0.764
ReReVST[65]	TIP 2020	0.171	0.271	0.200	**0.429**	2.951	0.863

注:粗体表示最优值,下划线表示次优值。"i"表示帧间隔。↑/↓:表示指标的衡量方向,"↑"表示该项指标的分数越高越好,"↓"表示该项指标的分数越低越好。+/−:表示指标分数的相对增加(+)或减少(−)。Δ 基线[9]:将本文方法应用至基线模型 CAP-VST[9]后,指标分数的相对增加(+)或减少(−)。ΔUniST[73]:相较 SOTA 模型 UniST[73],本文方法的指标分数的相对增加(+)或减少(−)。

与现有方法不同,本文方法充分利用了预训练大模型的丰富知识与能力,仅在训练阶段调用 Transformer,并通过结构损失与知识蒸馏的方式巧妙建立其与基线模型的联系,在不影响模型实时性能的情况下增强了风格化效果。与 S2WAT[68]相比,本文框架的推理参数量降至其 1/14.9,训练参数量降至其 1/5.26,推理时间加快了 3.95 倍,同时风格化效果具有优势(见表 1 与图 9),体现了本文方法的优越性。

4.4　消融实验

4.4.1　空间结构损失与频域纹理损失

本文在图 12 和表 4 中展示了消融研究结果以验证用于训练本文网络框架的每个损失项的有效性。本文方法的基线是 CAP-VST[9],采用作者发布的代码及其默认配置训练。如图 12 第 3 列所示,基线结果丢失了图像的很多内容细节,内容保真度低(如第 1 行中熊的面部以及第 2 行中钟楼的小表盘)。当加上基于拉普拉斯算子的空间结构损失

表 3 不同图像风格迁移方法的模型复杂度与效率分析

方法	推理阶段			训练阶段	
	参数量(Params)(M)	计算量(FLOPs)(G)	时间(Time)(s)	总参数量(M)	可训练参数量(M)
本文方法	4.09	79.95	0.059	99.70	10.38
Δ 基线[9]	0.00	0.00	0.000	+92.10	+6.29
ΔS2WAT[68]	−60.85 (14.9×)	−740.20 (9.26×)	−0.233 (3.95×)	+21.79	−54.58 (5.26×)
S2WAT[68]	64.94	820.15	0.292	77.91	64.96
AesFA[69]	3.22	50.03	0.010	6.73	3.22
CAP-VST[9]	4.09	79.95	0.059	7.60	4.09
TSSAT[28]	7.01	189.83	0.214	7.01	3.51
MicroAST[31]	0.47	11.06	0.004	3.98	0.47
StyTr2[70]	35.73	896.82	0.403	48.34	35.39
CCPL[30]	8.68	196.20	0.025	9.11	5.61
EFDM[25]	7.01	189.83	0.027	7.01	3.51
StyleFormer[71]	19.91	348.51	0.036	19.91	16.99
ArtFlow-W[10]	6.42	229.78	0.186	9.97	6.46
ArtFlow-A[10]	6.42	229.78	0.186	9.97	6.46
AdaAttN[27]	26.57	103.86	0.055	26.57	13.63
IECST[29]	20.91	267.74	0.034	29.41	16.46
CKD[7]	2.15	39.52	0.316	55.12	2.42

注:所有方法均在单张 NVIDIA Tesla V100(32GB) GPU 上测试;推理测试时图像的分辨率为 512×512。Δ 基线[9]:将本文方法应用至基线模型 CAP-VST[9]后,模型复杂度的相对增加(+)或减少(−)。ΔS2WAT[68]:相较 Transformer 模型 S2WAT[68],本文方法模型复杂度的相对增加(+)或减少(−)。

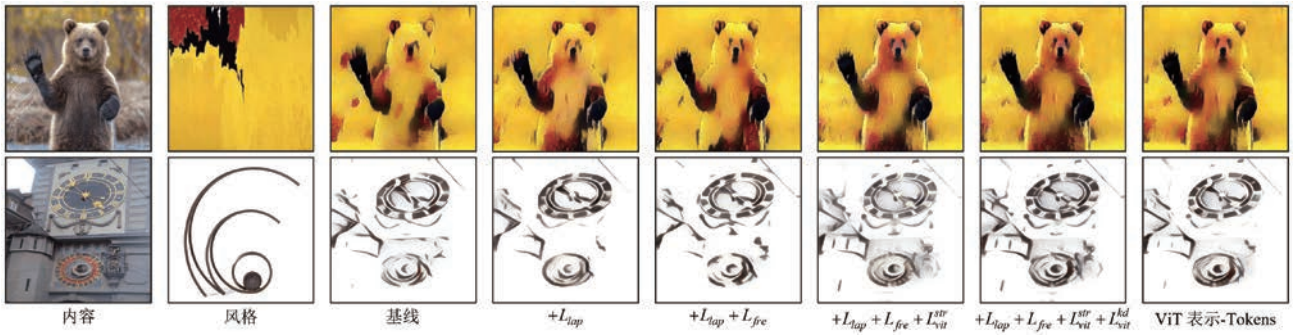


图 12 本文方法关键设计的消融实验可视化结果

表 4 本文方法关键设计的消融实验定量结果

方法	SSIM↑	LPIPS↓	L_c↓	L_s↓
基线[9]	0.528	0.696	4.383	0.523
$+L_{lap}$	0.561	0.678	4.156	0.555
$+L_{lap}+L_{fre}$	0.574	0.667	3.989	0.574
$+L_{lap}+L_{fre}+L_{vit}^{str}$ (Keys, $\lambda_{vit}^{str}=1$)	0.585	0.643	3.729	0.590
$+L_{lap}+L_{fre}+L_{vit}^{str}$ (Keys, $\lambda_{vit}^{str}=5$)	**0.610**	**0.599**	**3.284**	**0.640**
$+L_{lap}+L_{fre}+L_{vit}^{str}$ (Keys, $\lambda_{vit}^{str}=10$)	0.631	0.566	2.775	0.695
$+L_{lap}+L_{fre}+L_{vit}^{str}+L_{vit}^{kd}$ (Keys, $\lambda_{vit}^{str}=5$)	**0.607**	**0.600**	**3.348**	**0.628**
$+L_{lap}+L_{fre}+L_{vit}^{str}+L_{vit}^{kd}$ (Tokens $\lambda_{vit}^{str}=5$)	0.605	0.624	3.537	0.613

注:粗体表示最后所确定的超参数/特征表示的版本。↑/↓ 表示指标的衡量方向,↑表示该项指标的分数越高越好,↓表示该项指标的分数越低越好。

L_{lap} 后,如图 12 第 4 列所示,对象的边缘轮廓及其立体感等结构信息得到了增强(如第 1 行中熊的眼睛、耳朵以及第 2 行中钟楼的小表盘轮廓)。如图 12 第 5 列所示,频域纹理损失 L_{fre} 的加入进一步约束了图像中的细粒度信息,缓解了内容细节丢失的现象(如第 2 行中钟楼的表面纹理)。表 4 数据也进一步证明了以上结论。

4.4.2 ViT 语义结构损失

本节及下一节将讨论基于 DINO-ViT 指导的语义损失与知识蒸馏损失的有效性。如图 12 第 6 列所示,DINO-ViT 指导的语义结构损失 L_{vit}^{str} 显著增强了模型对内容源的物体结构及不同语义区域的感知能力,使得生成结果具有了立体感与景深,同时内容纹理也得到了进一步保留(如第 1 行中熊的躯

体及第 2 行中钟楼的表面纹理)。

此外,如表 4 第 5 行-第 7 行所示,本文还探索了 ViT 语义结构损失的约束强度(即权重超参数分别为 $\lambda_{vit}^{str}=1$, 5, 10 时)对风格迁移效果的影响。从表 4 中可以观察到,随着 λ_{vit}^{str} 的增大,内容结构保持度(SSIM、LPIPS 和 Content Loss)持续上升,但风格损失(Style Loss)也进一步加剧。由于本文希望在内容信息保持和艺术风格化之间取得较优的平衡,而不是只偏向某一边,因此最终选择了 $\lambda_{vit}^{str}=5$(表 4 中加粗表示)作为 ViT 语义结构损失的权重参数。

4.4.3　ViT 知识蒸馏损失

如图 12 第 6 列所示,尽管 ViT 语义结构损失 L_{vit}^{str} 效果显著,但其在一定程度上降低了风格化的笔触强度(第 2 行中钟楼的纹理与轮廓),同时也可能导致结果的风格化不一致(第 1 行中熊的面部)。而 ViT 知识蒸馏损失 L_{vit}^{kd} 的加入弥补了这一缺陷:通过强制编码器模仿 DINO 的特征空间,学习更通用和更强大的 DINO 知识,提升了编码器的特征表示能力。如图 12 及表 4 所示,相较于只有 L_{vit}^{str} 的情况,加入 L_{vit}^{kd} 能够让模型在维持内容结构保真度(SSIM、LPIPS 和 Content Loss)几乎不下降的前提下,显著提升风格化效果(Style Loss),证明了 L_{vit}^{kd} 的有效性。

4.4.4　ViT 特征表示的选择

根据 3.2.1 节的分析,在 DINO-ViT 的深层特征中,Keys 和 Tokens 都能够表示图像的内容结构信息。但如图 3 所示,相较于 DINO 的 Tokens 特征,Keys 特征捕获的内容结构及对象边界更加清晰(如第 1 行结果的松鼠轮廓),语义分割更加明确(如第 3 行结果的面部五官),也能抑制图像风格化后的失真伪影(如第 5 行结果的背景区域)。同时,如表 4 第 8 行-第 9 行所示,采用 Keys 作为 ViT 特征表示的内容指标优于采用 Tokens。因此,本文最终选用了从 DINO 最深层($L=12$)的注意力模块中提取的 Keys 作为 ViT 特征表示(表 4 中加粗表示)。

4.4.5　预训练 ViT 模型的选择

本文在 3.2.1 节分析了不同预训练 ViT 模型的特性与差异,并结合本文任务选择了 DINO[14] 构建 ViT 语义结构损失和知识蒸馏损失。为了进一步证明 DINO 和本任务的适配性,本节进行了预训练 ViT 模型选择的消融实验。作为对比,我们选择了目前主流的前沿预训练 ViT 模型(DINOv2[85]、CLIP[15]、原始 ViT[13])以构建 ViT 语义指导和知识蒸馏损失,并在保持其他配置相同的条件下重新训练模型。为了公平对比并突出性能,对于所有模型我们均采用 ViT/B 架构,同时根据各个模型的设计,选择各自 patch size 最小的架构(对于同一模型,patch size 越小,性能越优)。具体来说,我们选择了 DINO ViT/B-8、DINOv2 ViT/B-14、CLIP ViT/B-16、原始 ViT ViT/B-16 架构的预训练模型。实验结果如表 5 所示,可以看到,有监督 ViT 模型的 CLIP 与原始 ViT 在内容维持(除 SSIM 外)、风格匹配以及综合视觉质量三方面的得分均落后于自监督 ViT 模型的 DINO 系列。与我们在 3.2.1 节的分析相同,相较仅提供一个全局监督信号(标签)进行训练的有监督模型,自监督模型通过图像上下文信息自适应建模细粒度任务目标,在图像场景理解、内容语义感知等方面拥有更强大的特征提取能力。此外,DINOv2 在 DINO 的基础上改进了模型架构和训练策略,进一步提升了风格匹配能力,但在内容维持方面的权衡较多,最终导致综合视觉质量下降。相反,DINO 在内容和风格两方面的评分均位于前列(第 1 或第 2),通过优秀的内容-风格平衡获得了最优视觉质量评分。本文认为 DINOv1 和 v2 的差异源于它们各自不同的架构设计:DINO 采取了 patch size 为 8 的设计,而 DINOv2 为了提高训练效率将 patch size 设计为 14,这可能在一定程度上损害了对图像局部内容的细粒度建模能力。综上所述,本文最终选择了 DINO 作为预训练 ViT 模型来构建损失函数。

表 5　预训练 ViT 模型选择的消融实验

模型	SSIM ↑	LPIPS ↓	Content Loss ↓	Style Loss ↓	CLIP-IQA ↑
DINO[14]	<u>0.607</u>	**0.600**	**3.348**	<u>0.628</u>	**0.386**
DINOv2[85]	0.587	0.649	3.792	**0.601**	<u>0.378</u>
CLIP[15]	**0.611**	0.618	3.653	0.663	0.363
ViT[13]	0.606	<u>0.608</u>	<u>3.360</u>	0.672	0.368

注:粗体表示最优值,下划线表示次优值。↑/↓表示指标的衡量方向,↑表示该项指标的分数越高越好,↓表示该项指标的分数越低越好。

4.4.6　边缘滤波算子的选择

针对空间结构损失,本文进行了不同边缘滤波算子的消融实验。除了本文方法所用的 Laplace 算子[16]以外,我们还选择了主流的边缘滤波算子作为对比,分别为 8 邻域 Laplace 算子[86]、Sobel 算子[87]、CTSS 算子[88]、Gabor 算子[89]以及 Canny 算子[90]。其中,8 邻域 Laplace 算子[86]也是一种常用的 3×3 离散 Laplace 卷积核,其中心值为 −8,周围

8 个邻域值为 1。CTSS 算子[88]来源于最近的一种图像卡通化方法[88]，该方法提出的 CTSS 算子是一种改进 Sobel 算子，由引导滤波、恒定核卷积、归一化及高频滤波等过程组成，能够提取显著的卡通纹理边缘。实验结果如表 6 所示。可以看到，尽管 8 邻域 Laplace 算子、Sobel 算子以及 Gabor 算子在内容指标上都接近或超越了本文采用的 Laplace 算子，但全都进一步加重了风格化失真(Style Loss)。例如，Sobel 算子以严重的风格失真为代价换取了对内容结构的保持，最终导致结果的视觉质量(CLIP-IQA)下降。与此类似，Canny 算子的内容-风格平衡也表现不佳，影响了结果的视觉质量评分。需要强调的是，本文选用边缘滤波算子的目的是在尽量不影响风格化效果的情况下加强对内容结构的保持，即保持良好的内容-风格平衡，以获得综合最优的结果。因此本文最终选择了风格化评分(Style Loss)第二、视觉质量评分(CLIP-IQA)第一的 Laplace 算子。

表 6 边缘滤波算子选择的消融实验

算子	SSIM ↑	LPIPS ↓	Content Loss ↓	Style Loss ↓	CLIP-IQA ↑
Laplace[16]	0.607	0.600	3.348	0.628	**0.386**
Laplace[86] (8 邻域)	0.628	0.582	3.127	0.691	0.373
Sobel[87]	0.638	**0.574**	**3.007**	0.698	0.380
CTSS[88]	0.602	0.601	3.346	0.641	0.374
Gabor[89]	**0.654**	0.595	3.203	0.677	0.365
Canny[90]	0.587	0.612	3.453	**0.605**	0.367

注：粗体表示最优值，实下划线表示次优值，虚下划线表示最差值。↑/↓ 表示指标的衡量方向，↑表示该项指标的分数越高越好，↓表示该项指标的分数越低越好。

4.4.7 边缘滤波算子的卷积核大小

如表 7 和图 13 所示，本节探索了边缘滤波算子的卷积核大小对风格迁移效果的影响。具体来说，我们将卷积核大小分别设置为 3×3、5×5、7×7 以及 9×9 并进行实验。注意，由于离散 Laplace 卷积核[16]的常规尺寸为 3×3，因此对于 5×5、7×7 及 9×9 等扩展卷积核，我们采用了对应的 Laplacian and Gaussian(LoG)算子[16]。LoG 算子是结合了高斯平滑的 Laplace 算子，它有一个基于高斯函数的加权项，因此卷积核往往为 5×5 或 7×7 等较大尺寸。如图 13 第 1 行所示，当卷积核较小时(3×3)，提取的边缘图着重在对象边界和物体轮廓上，适当地保留了内容结构和场景布局而没有引入背景噪声；相反，随着卷积核的扩大(5×5 到 9×9)，边缘图的像素强度显著增加，平滑背景中出现了许多不应考虑的噪声纹理，失去了通过边缘图区分前景与背景、保持主要物体轮廓的作用。相应的，如图 13 第 2-4 行所示，当卷积核较小时(3×3)，结果图像能够在保留内容结构的同时充分表达艺术效果；而随着卷积核的扩大(5×5 到 9×9)，结果图像丢失了绝大部分风格化效果，破坏了内容-风格平衡。表 7 的数据也符合这一趋势。因此本文最终选择了常规尺寸的 3×3 Laplace 卷积核。

表 7 边缘滤波算子卷积核大小的消融实验

算子	SSIM ↑	LPIPS ↓	Content Loss ↓	Style Loss ↓
Laplace 3×3	0.607	0.600	3.348	**0.628**
ΔLoG 9×9	−0.120	+0.116	+1.833	−0.965
LoG 5×5	0.628	0.570	2.848	0.804
LoG 7×7	0.649	0.518	1.864	1.344
LoG 9×9	**0.727**	**0.484**	**1.515**	1.593

注：粗体表示最优值。↑/↓ 表示指标的衡量方向，↑表示该项指标的分数越高越好，↓表示该项指标的分数越低越好。ΔLoG 9×9：相较于 LoG 9×9，本文方法指标分数的相对增加(+)或减少(−)。

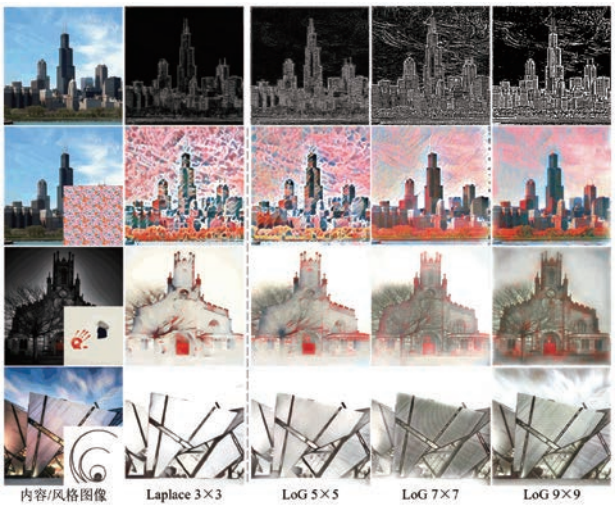


图 13 边缘滤波算子卷积核大小的消融实验

4.4.8 小波变换的分解尺度

针对频域纹理损失，本文探索了小波变换分解尺度的超参数设置。由于小波变换[22]具有多分辨率特性，因此它可以进行多级迭代分解以提取多尺度频域特征。具体来说，在每次分解之后，我们对低频部分(LL 子带)进行次级分解，分别完成了一至三级分解以构造频域损失。实验结果如表 8 所示，随着分解尺度的加深(一级到三级)，风格化图像呈现内容保持度上升、风格化效果与视觉质量下降的趋势，并且风格评分的下降幅度较大。对此，本文推测是由于多级分解加强了对图像低频信息的挖掘，

从而引入了内容图像的风格(低频)信息,导致对风格图像的学习不够深入。考虑到本文引入小波变换的初衷是为了提取内容图像的细粒度纹理(高频)信息,而非风格(低频)信息,因此我们最终采用了风格化效果与视觉质量最优的一级小波分解。

表 8　小波变换分解尺度的消融实验

分解尺度	SSIM ↑	LPIPS ↓	Content Loss ↓	Style Loss ↓	CLIP-IQA ↑
一级	0.607	0.600	3.348	**0.628**	**0.386**
二级	0.610	0.593	3.235	0.648	0.365
三级	**0.625**	**0.575**	**3.001**	0.696	0.359

注:粗体表示最优值。↑/↓表示指标的衡量方向,"↑"表示该项指标的分数越高越好,"↓"表示该项指标的分数越低越好。

4.5　本文方法在不同骨干网络上的应用

本文方法作为一种独立的训练框架,可以即插即用于大多数任意风格迁移模型。除了 4.2 节的基线模型 CAP-VST[9],我们还将本文方法应用到以下骨干网络:AesFA[69]、EFDM[25]、ArtFlow[10]。对比结果如图 14 和表 9 所示。AesFA 为了追求实时性能,采用了简单的网络结构并引入八度卷积以压缩参数量,模型的特征表示能力不可避免地下降,导致同一语义区域的风格化不一致(如第 1 组的面部)。EFDM 设计了精准特征分布匹配算法,强调对高阶风格特征分布的转换,忽略了对内容特征的维持,导致结果的内容语义失真严重(如第 4 组的熊)。ArtFlow 中级联的可逆转换模块和压缩模块会使特征通道数呈指数级增加,导致在前向推理过程中积累冗余信息[9],从而对风格化产生负面影响(如第 2 组的天空)。

相反,本文方法的加入显著改善了以上三种模型的生成质量。通过本文框架的重新训练,所有增强模型都可以生成具有平滑背景(如第 2 组天空)和完整内容结构(如第 4 组的熊)的风格化图像,同时

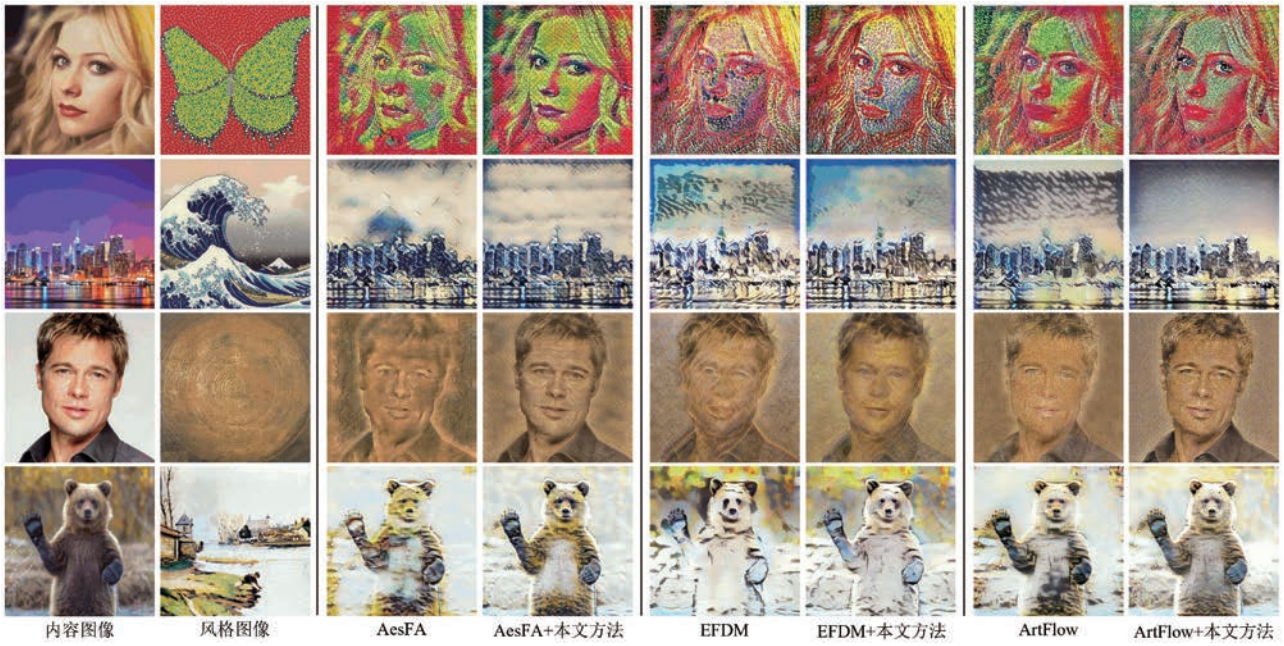


图 14　本文方法在不同骨干网络(AesFA[69]、EFDM[25]、ArtFlow[10])上的应用

表 9　本文方法在不同骨干网络上应用的定量结果

方法	SSIM ↑	LPIPS ↓	Content Loss ↓	Style Loss ↓	CLIP-IQA ↑	SSIM-CLIP ↓
AesFA[69]	0.407	0.686	4.209	**0.587**	0.341	2.123
AesFA[69]+本文方法	**0.542**	**0.578**	**2.965**	0.777	**0.349**	**2.036**
EFDM[25]	0.312	0.710	4.792	**0.879**	0.369	2.281
EFDM[25]+本文方法	**0.351**	**0.665**	**4.031**	0.894	**0.371**	**2.250**
ArtFlow[10]	0.439	0.614	3.735	**0.737**	0.363	2.132
ArtFlow[10]+本文方法	**0.518**	**0.513**	**3.727**	0.825	**0.374**	**2.083**
SANet[26]	0.396	0.581	3.290	**0.641**	**0.367**	2.202
SANet[26]+IECST[29]	0.425	0.546	3.228	0.995	0.364	**2.081**
SANet[26]+本文方法	**0.482**	**0.501**	**2.684**	0.833	0.333	2.110

注:粗体表示同一骨干网络上指标的较优值。↑/↓表示指标的衡量方向,↑表示该项指标的分数越高越好,↓表示该项指标的分数越低越好。

区域风格化不一致的现象也得到了显著改善(如第1组的面部)。同时,如表9所示,在本文框架的加持下,所有方法的内容保持度、视觉质量(CLIP-IQA)和内容-风格平衡(SSIM-CLIP)评分都获得了提升,进一步证明了本文方法的普适性。最后需要强调的是,本文方法并不影响推理阶段的模型复杂度和运行效率,这意味着经过本文方法重新训练后,AesFA等模型可以实现视觉质量更好的实时超分辨率风格迁移。

4.6 与现有方法损失函数的对比实验

为了进一步证明本文所提出损失函数的优势,我们进行了与现有方法损失函数的对比实验,如表9和图15所示。具体来说,我们选用SANet[26]作为基线模型,该方法提出了风格注意力网络(Style-Attention Network),相较传统方法能够进一步理解和表现目标风格的细节,避免风格信息丢失。SANet采用的损失函数有内容损失 L_c 、风格损失 L_s 及恒等损失 L_{identity} ,均为风格迁移任务中基础的损失函数。在此基础上,我们对比了IECST[29]方法。该方法将SANet作为骨干网络,引入对比学习[32]和生成对抗策略[80]设计了一系列损失函数。具体来说,IECST方法由三个损失函数构成:对抗损失 L_{adv} 、内容对比学习损失 $L_{\text{c-contra}}$ 、风格对比学习损失 $L_{\text{s-contra}}$ 。作为对比,我们也将SANet作为骨干网络,添加本文所提出损失函数后重新训练。训练时,三种模型均按照IECST的参数从零开始训练160000次,且除损失函数外所有条件不变。实验结果如表9和图15所示。从图15中可以看到,由于SANet[26]设计的风格注意力网络是通过交叉注意力直接调制内容和风格特征,模块较为简单,因此会导致内容结构的扭曲失真(如第2行和第5行的结果图像),同时会将风格图像的内容信息引入到结果图像中,产生不和谐伪影(如第3行和第4行,结果图像的天空/面部出现来自风格图像中人物的眼睛样伪影)。在IECST[29]作为损失函数加入后,得益于对比学习和生成对抗策略,结果图像的视觉质量显著提高,质感更加类似于真实艺术品。但同时由于GAN的外部学习特性,使得模型过拟合于判别器特征,无法完全遵循单一参考风格图像的风格信息,导致结果图像的整体风格(如色调、笔触、纹理)时常偏离参考风格图像(如第1行与第2行的结果图像,以及表9的Style Loss),同时会在某种程度上改变内容图像的语义结构(如第4行和第5行,结果图像的人脸相较内容图像发生了一定变化)。而在本文框架下重新训练的SANet,不仅消除了来自风格图像的内容伪影,还保持了内容主体的语义结构,同时避免了结果的整体风格偏离风格图像。此外,如表9所示,加入本文损失函数后,风格化结果取得了最优的内容维持(SSIM、LPIPS、Content Loss)评分,证明了本文方法相较其他方法的优越性。

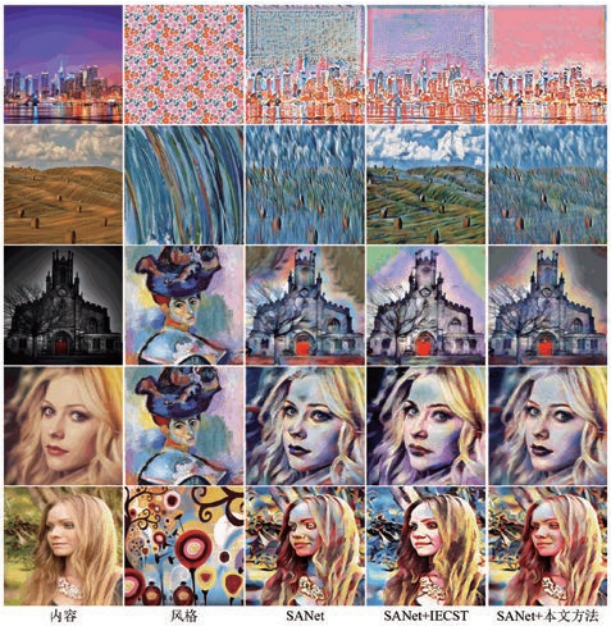


图15 与现有方法损失函数的对比实验

针对以上结果,我们认为本文方法的独特优势主要有以下三点:

首先,本文在深入分析后采用了具有空间感知和语义定位特性的自监督预训练ViT模型DINO[14]。相较VGG[8]等传统CNN模型而言,DINO具备长距离依赖建模能力,可以更好地建模全局内容语义;相较CLIP[15]等有监督ViT模型而言,DINO对于涉及图像内容的场景布局和对象边界等属性识别度更高,有助于维持内容保真度和一致性风格化。

其次,相较现有从零开始训练ViT模型的Transformer类方法[68,70-71],本文方法仅将ViT模型作为增强网络放在训练阶段。该策略不仅大大减少了推理阶段的模型复杂度和计算开销,在提升性能的同时保持了推理效率,还能够作为训练插件推广至大多数风格迁移方法中。特别是对于那些网络架构简单以追求实时性能的模型(如AesFA[69]),本文方法可以无负担地提升其实时推理性能,以更好地迁移至边缘设备上。

最后,相较于同样在训练阶段添加辅助模型(判别器)的IECST[29]而言,本文方法选择了ViT而非

GAN,一方面避免了外部学习带来的负面影响,另一方面避免了 GAN 训练的不稳定性和复杂性,减少了训练阶段的调参难度,使本文方法更容易推广至其他模型。

综上所述,本文方法作为一种独立的网络框架,具备性能提升显著、兼容性高、不影响实时性能等优势。

4.7　本文方法在图像翻译任务上的应用

本文方法的内容跨域一致性保持能力不仅适用于风格迁移任务,还可以很好地迁移至其他图像生成任务中,如图像翻译。本节以图像翻译的经典方法 CUT[91] 为例,展示本文方法的良好迁移性。首先,我们将本文损失函数添加到 CUT 中,然后使用该任务常用的 horse2zebra 数据集进行训练,所有训练参数与 CUT 默认值相同。实验结果如图 16 所示,CUT 无法准确感知图像的内容语义与场景布局,时常错误地将非目标对象(如左边第 2 行的背景、右边第 1 行的人群)转换至“斑马”域,同时也会在平滑背景中添加杂乱的纹理伪影(如左边第 1 行、右边第 1-3 行的结果),使结果图像的视觉质量下降。相反,应用本文方法后的 CUT 不仅能够正确转换目标对象(马→斑马)的外观,还避免了前景人群和背景天空等非目标语义的外观发生改变,同时消除了不和谐的杂乱纹理,有效提升了结果图像的整体视觉质量,证明了本文方法良好的可迁移性和有效性。

图 16　本文方法在图像翻译任务上(CUT[91])的应用

4.8　局限性

尽管 ViT 语义指导与结构增强损失实现了高内容保真度的风格化,但本文方法仍旧存在局限性。首先,当内容图像与风格图像的固有色彩差异过大时,本文方法可能会无法完全消除来自内容源的影响,导致风格化结果中残留内容图像的色彩,如图 17(a)所示。这可能是由于本文方法对于内容的约束太强,且无法充分解耦内容的结构与风格导致的。此外,当风格纹理过于复杂且风格语义与内容语义相关性较低时,本文方法可能无法准确对应风格与内容的语义,导致结果中局部语义的风格化不正确,如图 17(b)所示。这可能是由于本工作采用的转换模块为 cWCT[9],一种匹配全局统计分布的模块,其对于局部语义的对齐能力较差。

我们的未来工作是解决这些局限性。例如,结合多种 ViT 大模型,在 DINO[14] 的内容保持基础上引入 CLIP[15] 以增强风格化效果;其次,通过基于注意力的转换模块实现更加细粒度的风格化等。

图 17　本文方法的局限性

4.9　实验总结

本节通过图像风格迁移的定性比较,证明了本文方法能够保持全局内容结构的完整性与局部语义区域的风格化一致性,同时借助边缘图与深度图可视化实验,进一步证明了本文方法可以在有效保持内容层次与景深的基础上生成富有艺术表现力的风格化图像。此外,本节通过视频风格迁移的定性比较,证明了本文方法能够在避免艺术效果过度损失的基础上提升基线模型[9] 的运动帧稳定性和内容语义保持度,展示了本文方法的优越性。

在图像风格迁移的定量比较中,本文方法不仅在内容保持方面取得了显著优势(最优 SSIM 值,次优 LPIPS、Content Loss 值),还维持了与 Trans-

former 类方法[68] 相当的风格化效果(Style Loss 值),最终达到了综合最优的视觉质量(最优 CLIP-IQA、SSIM-CLIP 值),证明了本文方法的有效性与优越性。在视频风格迁移的定量比较中,本文方法在时间一致性(Temporal Loss、LPIPS)、内容保持度(Content Loss)及风格化效果(Style Loss)三个方面均取得了最优或次优值,证明了本文方法的有效性与平衡性。此外,通过模型复杂度与效率分析,证明了本文方法可以作为训练插件兼容于现有实时风格迁移模型的有效性与灵活性,以及本文方法相较于现有 Transformer 类方法[68,70] 在训练成本和推理效率上的优越性。

在消融实验部分,本节分别证明了本文设计的空间结构损失、频域纹理损失、ViT 语义结构损失、ViT 知识蒸馏损失的有效性。同时,证明了本文方法的 ViT 特征表示选择、预训练 ViT 模型选择、边缘滤波算子选择、边缘滤波算子卷积核大小、小波变换分解尺度等各类模块选择及超参数设置的合理性与必要性。

进一步,通过本文方法在不同骨干网络上的应用实验,证明了本文方法能够从内容保持(SSIM、LPIPS、Content Loss)、视觉质量(CLIP-IQA)、内容-风格平衡(SSIM-CLIP)三方面改进现有方法的生成效果,表明了本文方法的有效性与普适性。同时,通过与现有方法损失函数的对比实验,证明了本文方法相较其他 SOTA 损失函数设计的优越性。此外,通过在图像翻译任务上的应用实验,证明了本文方法的内容跨域一致性保持能力不仅适用于风格迁移任务,还可以很好地迁移至其他图像生成任务中。

最后,本文方法存在内容图像色彩残留与局部语义风格化不正确等局限性。对此,计划在未来工作中结合多种 ViT 大模型与基于注意力的转换模块,进一步探索艺术风格迁移方法的设计与应用。

5 结 论

本文提出了一种基于 ViT 指导的语义结构损失和知识蒸馏损失,通过 DINO 强大的语义表征能力保持内容结构的跨域一致性,并将丰富且多样的 DINO 知识提炼到编码器,显著缓解了内容保真度低、风格化不一致的问题。为了进一步提升模型的结构与纹理感知能力,本文还设计了一种基于拉普拉斯算子的结构损失函数,同时利用小波变换生成的频域表示鼓励模型捕获高频纹理信号,确保了风格化图像的全局连贯性。实验表明,本文方法在内容结构保持和局部语义区域风格化一致性方面显著优于现有方法,并且可以无缝集成到现有方法中,证明了所提方法的有效性和灵活性。未来将进一步探索更多预训练大模型对小模型训练的影响,并整合至现有算法中以改善生成质量。同时考虑将现有的二维风格迁移方法扩展至更多具有现实需求的场景中,如数字三维场景[92-93] 等。

参 考 文 献

[1] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2414-2423

[2] Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution//Proceedings of the Computer Vision-ECCV 2016: 14th European Conference. Amsterdam, The Netherlands, 2016: 694-711

[3] Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1510-1519

[4] Li Y, Fang C, Yang J, et al. Universal style transfer via feature transforms//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 385-395

[5] Park D Y, Lee K H. Arbitrary style transfer with style-attentional networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5873-5881

[6] Wang Z, Zhao L, Chen H, et al. Diversified arbitrary style transfer via deep feature perturbation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 7786-7795

[7] Wang H, Li Y, Wang Y, et al. Collaborative distillation for ultra-resolution universal style transfer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 1857-1866

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014

[9] Wen L, Gao C, Zou C. CAP-VSTNet: Content affinity preserved versatile style transfer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 18300-18309

[10] An J, Huang S, Song Y, et al. ArtFlow: Unbiased image style transfer via reversible neural flows//Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 862-871

[11] Kingma D P, Dhariwal P. Glow: Generative flow with invertible 1x1 convolutions//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 10236-10245

[12] Tumanyan N, Bar-Tal O, Bagon S, et al. Splicing vit features for semantic appearance transfer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 10738-10747

[13] Alexey D, Lucas B, Alexander K, et al. An image is worth 16x16 words: Transformers for image recognition at scale//Proceedings of the International Conference on Learning Representations. Virtual, 2021: 1-12

[14] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 9630-9640

[15] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual, 2021: 8748-8763

[16] Van Vliet L J, Young I T, Beckers G L. A nonlinear Laplace operator as edge detector in noisy images. Computer Vision, Graphics, and Image Processing, 1989, 45(2): 167-195

[17] Yang M, Wang Z, Feng W, et al. Improving few-shot image generation by structural discrimination and textural modulation//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada, 2023: 7837-7848

[18] Gao Y, Wei F, Bao J, et al. High-fidelity and arbitrary face editing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 16110-16119

[19] Schwarz K, Liao Y, Geiger A. On the frequency bias of generative models//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual, 2021, 34: 18126-18136

[20] Xu Z Q J, Zhang Y, Luo T, et al. Frequency principle: Fourier analysis sheds light on deep neural networks. Communications in Computational Physics, 2020, 28 (5): 1746-1767

[21] Yang M, Wang Z, Chi Z, Zhang Y. FreGAN: Exploiting frequency components for training gans under limited data//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 33387-33399

[22] Daubechies I. The wavelet transform, time-frequency localization and signal analysis. IEEE Transactions on Information Theory, 1990, 36(5): 961-1005

[23] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context//Proceedings of the Computer Vision-ECCV 2014: 13th European Conference. Zurich, Switzerland, 2014: 740-755

[24] Karayev S, Trentacoste M, Han H, et al. Recognizing image style//Proceedings of the British Machine Vision Conference. Nottingham, UK, 2014: 1-11

[25] Zhang Y, Li M, Li R, et al. Exact feature distribution matching for arbitrary style transfer and domain generalization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 8025-8035

[26] Park D Y, Lee K H. Arbitrary style transfer with style-attentional networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5873-5881

[27] Liu S, Lin T, He D, et al. Adaattn: Revisit attention mechanism in arbitrary neural style transfer//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 6629-6638

[28] Chen H, Zhao L, Li J, et al. Tssat: Two-stage statistics-aware transformation for artistic style transfer//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada, 2023: 6878-6887

[29] Chen H, Wang Z, Zhang H, et al. Artistic style transfer with internal-external learning and contrastive learning//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual, 2021, 34: 26561-26573

[30] Wu Z, Zhu Z, Du J, et al. Ccpl: Contrastive coherence preserving loss for versatile style transfer//Proceedings of the Computer Vision-ECCV 2022: 17th European Conference. Tel Aviv, Israel, 2022: 189-206

[31] Wang Z, Zhao L, Zuo Z, et al. Microast: Towards superfast ultra-resolution arbitrary style transfer//Proceedings of the AAAI Conference on Artificial Intelligence. Washington DC, USA, 2023, 37(3): 2742-2750

[32] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations//Proceedings of the International Conference on Machine Learning. Virtual, 2020: 1597-1607

[33] Amir S, Gandelsman Y, Bagon S, et al. On the effectiveness of vit features as local semantic descriptors//Proceedings of the Computer Vision-ECCV 2022 Workshops. Tel Aviv, Israel, 2022: 39-55

[34] Zhou Y, Chen Z, Huang H. Deformable one-shot face stylization via dino semantic guidance//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 7787-7796

[35] Buciluǎ C, Caruana R, Niculescu-Mizil A. Model compression//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 535-541

[36] Ba J, Caruana R. Do deep nets really need to be deep? //Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada, 2014:

2654-2662

[37] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015

[38] Cui K, Yu Y, Zhan F, et al. Kd-dlgan: Data limited image generation via knowledge distillation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 3872-3882

[39] Wang H, Li Y, Wang Y, et al. Collaborative distillation for ultra-resolution universal style transfer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 1857-1866

[40] Dong B, Jiang Q, Shen Z. Image restoration: Wavelet frame shrinkage, nonlinear evolution pdes, and beyond. Multiscale Modeling & Simulation, 2017, 15(1): 606-660

[41] Li B, Chen X. Wavelet-based numerical analysis: A review and classification. Finite Elements in Analysis and Design, 2014, 81: 14-31

[42] Phung H, Dao Q, Tran A. Wavelet diffusion models are fast and scalable image generators//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 10199-10208

[43] Yang M, Wang Z, Chi Z, et al. Wavegan: Frequency-aware gan for high-fidelity few-shot image generation//European Conference on Computer Vision. Tel Aviv, Israel, 2022: 1-17

[44] Li Z, Kuang Z S, Zhu Z L, et al. Wavelet-based texture reformation network for image super-resolution. IEEE Transactions on Image Processing, 2022, 31: 2647-2660

[45] Deng X, Yang R, Xu M, et al. Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 3076-3085

[46] Kim M W, Cho N I. Whfl: Wavelet-domain high frequency loss for sketch-to-image translation//Proceedings of the IEEE/CVF Winter Conference on applications of computer vision. Waikoloa, Hawaii, 2023: 744-754

[47] Ma L, Gao T, Jiang H, et al. Waveipt: Joint attention and flow alignment in the wavelet domain for pose transfer//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 7215-7225

[48] Yoo J, Uh Y, Chun S, et al. Photorealistic style transfer via wavelet transforms//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 9035-9044

[49] Cho H, Lee J, Chang S, et al. One-shot structure-aware stylized image synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 8302-8311

[50] Liu S, Zhu T. Structure-guided arbitrary style transfer for artistic image and video. IEEE Transactions on Multimedia, 2021, 24: 1299-1312

[51] Xie S, Tu Z. Holistically-nested edge detection//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1395-1403

[52] Wu J, Hou L, Li Z, et al. Preserving structural consistency in arbitrary artist and artwork style transfer//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(3): 2830-2838

[53] Gao X, Zhang Y, Tian Y. Learning to incorporate texture saliency adaptive attention to image cartoonization//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA, 2022: 7183-7207

[54] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2223-2232

[55] Kwon G, Ye J C. Clipstyler: Image style transfer with a single text condition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 18062-18071

[56] Yang S, Hwang H, Ye J C. Zero-shot contrastive loss for text-guided diffusion image style transfer//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 22873-22882

[57] Wysoczańska M, Siméoni O, Ramamonjisoa M, et al. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation//Proceedings of the Computer Vision-ECCV 2024. Milan, Italy, 2025: 320-337

[58] Jiang D, Liu Y, Liu S, et al. From clip to dino: Visual encoders shout in multi-modal large language models. arXiv preprint arXiv:2310.08825, 2023

[59] Jolliffe I T. Principal Component Analysis. 2nd Edition. New York: Springer New York, 2002

[60] Wang Z, Zhao L, Xing W. Stylediffusion: Controllable disentangled style transfer via diffusion models//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 7677-7689

[61] Zhang Y, Huang N, Tang F, et al. Inversion-based style transfer with diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 10146-10156

[62] Shechtman E, Irani M. Matching local self-similarities across images and videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8

[63] Kolkin N, Salavon J, Shakhnarovich G. Style transfer by relaxed optimal transport and self-similarity//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 10051-10060

[64] Levin A, Lischinski D, Weiss Y. A closed-form solution to natural image matting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(2): 228-242

[65] Wang W, Yang S, Xu J, et al. Consistent video style trans-

fer via relaxation and regularization. IEEE Transactions on Image Processing, 2020, 29: 9125-9139

[66] Butler D J, Wulff J, Stanley G B, et al. A naturalistic open source movie for optical flow evaluation//Computer Vision-ECCV 2012: 12th European Conference on Computer Vision. Florence, Italy, 2012: 611-625

[67] Kingma D P, Ba J. Adam: A method for stochastic optimization//Proceedings of the International Conference on Learning Representations. San Diega, USA, 2015: 1-13

[68] Zhang C, Xu X, Wang L, et al. S2wat: Image style transfer via hierarchical vision transformer using strips window attention//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024, 38(7): 7024-7032

[69] Kwon J, Kim S, Lin Y, et al. AesFA: An Aesthetic Feature-Aware Arbitrary Neural Style Transfer//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024, 38(12): 13310-13319

[70] Deng Y, Tang F, Dong W, et al. Stytr2: Image style transfer with transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 11326-11336

[71] Wu X, Hu Z, Sheng L, et al. Styleformer: Real-time arbitrary style transfer via parametric style composition//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 14618-14627

[72] Chen Y, Fan H, Xu B, et al. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 3434-3443

[73] Gu B, Fan H, Zhang L. Two birds, one stone: A unified framework for joint learning of image and video style transfers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 23545-23554

[74] Deng Y, Tang F, Dong W, et al. Arbitrary video style transfer via multi-channel correlation//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021, 35(2): 1210-1217

[75] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612

[76] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 586-595

[77] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1106-1114

[78] Wang J, Chan K C K, Loy C C. Exploring clip for assessing the look and feel of images//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(2): 2555-2563

[79] Wright M, Ommer B. Artfid: Quantitative evaluation of neural style transfer//DAGM German Conference on Pattern Recognition. Konstanz, Germany, 2022: 560-576

[80] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680

[81] Wang X, Chen H, Sun P, et al. AdvST: Generating Unrestricted Adversarial Images via Style Transfer. IEEE Transactions on Multimedia, 2024, 26: 4846-4858

[82] Soria X, Pomboza-Junez G, Sappa A D. Ldc: Lightweight dense cnn for edge detection. IEEE Access, 2022, 10: 68281-68290

[83] Ranftl R, Lasinger K, Hafner D, et al. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(3): 1623-1637

[84] Ke Z, Liu Y, Zhu L, et al. Neural preset for color style transfer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 14173-14182

[85] Oquab M, Darcet T, Moutakanni T, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023

[86] Gonzalez R C, Woods R E, Masters B R. Digital Image Processing. 3rd Edition. USA: Prentice-Hall, Inc., 2006

[87] Sobel I, Feldman G. A 3x3 isotropic gradient operator for image processing. The Stanford Artificial Intelligence Project, 1968: 271-272

[88] Gao X, Zhang Y, Tian Y. Learning to incorporate texture saliency adaptive attention to image cartoonization//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA, 2022: 7183-7207

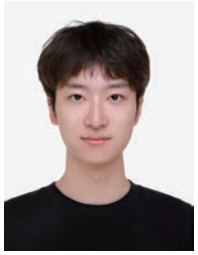
[89] Gabor D. Theory of communication. part 1: The analysis of information. Journal of the Institution of Electrical Engineers-part III: Radio and Communication Engineering, 1946, 93(26): 429-441

[90] Canny J. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986 (6): 679-698

[91] Park T, Efros A A, Zhang R, et al. Contrastive learning for unpaired image-to-image translation//Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, 2020: 319-345

[92] Zhang S K, Tam H, Li Y, et al. Scenedirector: Interactive scene synthesis by simultaneously editing multiple objects in real-time. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(8): 4558-4569

[93] Zhang S K, Li Y X, He Y, et al. Mageadd: Real-time interaction simulation for scene synthesis//Proceedings of the 29th ACM International Conference on Multimedia. Virtual, 2021: 965-973



PAN Shu-Yu, M. S. candidate. His research interests include computer vision and image style transfer.

ZHAO Zheng-Peng, M. S., Professor. His research interests include signal and information processing, computer systems and applications.

YANG Qiu-Xia, Ph. D. candidate. Her research interests include computer vision and image translation.

PU Yuan-Yuan, Ph. D., Professor. Her research interests include digital image processing and scientific understanding of visual arts.

GU Jin-Jing, Ph. D., Lecturer. Her research interests include cross-media semantic analysis and understanding, video abnormal behavior recognition.

XU Dan, Ph. D., Professor. Her research interests include graphics rendering and image fusion.

Background

This study focuses on the task of style transfer in the field of computer vision, specifically on artistic style transfer. Artistic style transfer has been a long-standing research hotspot in computer vision, aiming to transfer the artistic style of a reference style image to a content image while preserving the semantic structure of the content image.

In recent years, with the rise of deep learning, the task of style transfer has entered the era of neural style transfer and has seen rapid development. Current methods, through matching the global statistical distribution of features and local attention mechanisms, have been able to achieve style transfer effectively, generating results with good visual quality. However, existing deep learning-based artistic style transfer methods still face a major challenge: they cannot maintain the cross-domain semantic structure consistency from the content domain to the style domain well during the transfer process, resulting in low content fidelity and inconsistent stylization in the generated results.

To address the above issues, this paper proposes an artistic style transfer method based on ViT (Vision Transformer) semantic guidance and structure-aware enhancement. Firstly, a pre-trained DINO-ViT model is used to establish strong and consistent content structure representations in both the content and style domains, and two loss functions are designed: (1) a semantic structure loss based on DINO keys self-similarity to maintain cross-domain consistency of the content source; (2) a knowledge distillation loss in the DINO feature space to enhance the feature extraction capability of the encoder. To further enhance the model's structure awareness, a spatial structure loss based on the Laplacian operator and a frequency domain texture loss based on wavelet transform are proposed, enhancing constraints on edge contours and fine textures from both spatial and frequency domains. Qualitative and quantitative results on general datasets for style transfer tasks show that the proposed method can not only produce results with high content fidelity and consistent stylization but also be applied to existing methods to further improve the visual quality of generated results.

This work was supported by the National Natural Science Foundation of China under Grant 61271361, 61761046, and 62362070; in part by the Key Project of Applied Basic Research Program of Yunnan Provincial Department of Science and Technology under Grant 202001BB050043; in part by the Basic Research Project of Science and Technology Program of Yunnan Provincial Science and Technology Department under Grant 202401AS070149; in part by the Yunnan Key Laboratory of low-light Night Vision Technology and Intelligent Visual Navigation under Grant 202449CE340004; in part by the Postgraduate Science Foundation of Yunnan University under Grants ZC-23235984 and KC-23235986.