doi:10.3969/j.issn.1003-3114.2025.05.017
引用格式:张嫱,何家成,孔令军.面向艺术图像分类的频域增强对比学习方法研究[J].无线电通信技术,2025,51(5):1046-1055.[ZHANG Qiang,HE Jiacheng,KONG Lingjun. Research on Frequency Domain Enhanced Contrastive Learning Method for Art Image Classification[J]. Radio Communications Technology,2025,51(5): 1046-1055.]

面向艺术图像分类的频域增强对比学习方法研究

张　嫱[1], 何家成[2], 孔令军[3*]
(1.金陵科技学院 艺术学院,江苏 南京 211169;
2.南京邮电大学 计算机学院、软件学院、网络空间安全学院,江苏 南京 210023;
3.金陵科技学院 网络与通信工程学院,江苏 南京 211169)

摘　要:针对艺术图像分类任务中存在的样本稀缺、风格多样与纹理复杂等挑战,提出了一种新的自监督学习框架——频域掩码对比(Frequency-Masked Contrast,F-MaCo),以双分支对比学习为基础,通过二维离散小波变换(Discrete Wavelet Transform,DWT)将图像转换到频域,实现动态频域掩码增强。引入感知损失驱动的权重调整机制,有效捕捉艺术图像的多尺度特征和丰富的纹理信息。实验结果表明,F-MaCo 在 MAMe、Kaokore、Artbench10 和 ArtDL 四个艺术图像数据集上均取得了最优性能,Top-1 准确率分别达到 73.72%、77.38%、58.38%和 68.31%,验证了其在艺术图像表征学习任务中的有效性与鲁棒性。

关键词:艺术图像分类;自监督学习;小波变换;对比学习;频域掩码

中图分类号:TP391.4　　**文献标志码**:A　　**开放科学(资源服务)标识码(OSID)**:

文章编号:1003-3114(2025)05-1046-10

Research on Frequency Domain Enhanced Contrastive Learning Method for Art Image Classification

ZHANG Qiang[1], HE Jiacheng[2], KONG Lingjun[3*]
(1. College of Art, Jinling Institute of Technology, Nanjing 211169, China;
2. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
3. School of Network and Communication Engineering, Jinling Institute of Technology, Nanjing 211169, China)

Abstract: To address the challenges of data scarcity, stylistic diversity, and complex textures in art image classification, a novel self-supervised learning framework is proposed——Frequency-Masked Contrast (F-MaCo). Built upon a dual-branch contrastive learning paradigm, F-MaCo leverages a two-dimensional Discrete Wavelet Transform (DWT) to project images into the frequency domain, enabling dynamic frequency-domain masking augmentation. Additionally, a perceptual loss-driven weighting mechanism is introduced to effectively capture the multi-scale features and rich textures information of art images. Experimental results demonstrate that F-MaCo achieves state-of-the-art performance on four art image datasets—MAMe, Kaokore, Artbench10, and ArtDL—with Top-1 accuracies of 73.72%, 77.38%, 58.38%, and 68.31%, respectively, validating its effectiveness and robustness in art image representation learning.

Keywords: art image classification; self-supervised learning; wavelet transform; contrastive learning; frequency domain masking

收稿日期:2025-06-16
基金项目:江苏高校哲学社会科学研究项目(2025SJYB0417);江苏省研究生科研创新与实践计划(SJCX24_0324);江苏省高等学校基础科学(自然科学)研究重大项目 (22KJA510009);金陵科技学院高层次人才科研启动资金 (jit-b-202110)
Foundation Item: Philosophy and Social Sciences Research Project in Jiangsu Province's Universities(2025SJYB0417); Postgraduate Research & Practice Innovation Program of Jiangsu Province(SJCX24_0324); Major Project of Basic Science (Natural Science) Research in Higher Education Institutions in Jiangsu Province (22KJA510009); Jinling University of Science and Technology's High-level Talent Research Start-up Fund (jit-b-202110)

0 引言

随着人工智能技术的不断发展,计算机视觉在艺术领域的应用日益广泛。艺术图像分类作为计算机视觉的一个重要研究方向,在文化遗产保护、艺术风格分析、数字博物馆建设等领域具有重要应用价值。然而,艺术图像分类面临诸多挑战,如样本稀少导致模型训练数据不足,难以学习到足够的特征模式;风格多样性使得不同艺术风格的图像特征差异较大,增加了模型统一识别的难度;细节丰富则要求模型具备强大的特征捕捉能力,使得传统的监督学习方法难以取得理想效果。此外,已有一些研究专门关注艺术图像在风格识别与分类任务中的特征建模。例如,有学者提出利用色彩分布、笔触纹理和构图布局等艺术特征进行风格识别[1],也有研究尝试结合图像结构与美学特征用于艺术作品分类[2]。这些方法虽在一定程度上缓解了艺术图像分类中风格多样性带来的挑战,但大多仍依赖手工特征或传统模型,难以深入挖掘图像中的结构潜力。近年来,研究者开始尝试引入更具表达能力的表征方式,如将频域特征与图像空间信息相结合,以提升模型对艺术图像结构与纹理的理解能力。同时,自监督学习技术因其无需大量标注数据的优势,成为应对该问题的一种潜在方向。它是一种新的机器学习模式,允许模型通过从无标签数据本身构建监督信号进行学习[3]。近年来,研究者们从多种视角出发对自监督学习技术进行改进[4-6],根据其预训练任务的性质,可以将自监督学习技术分为两大类:对比学习与掩码学习。

在计算机视觉领域,对比学习的核心思想是使模型能够通过拉近相似图像特征的同时,推远不同图像特征从而学习到强大的视觉表征[7]。这种方法使模型能够在无标签情况下,通过自监督从数据中学习有意义的特征。SimCLR[8]是一个简单而有效的对比学习框架,它通过数据增强生成相似样本对(正样本对),并使用线性投影头将图像嵌入投射到高维空间。SimCLR 的关键在于使用强数据增强(如颜色抖动、裁剪、旋转等)和大批量训练,使其在自监督学习的视觉任务中取得了优异成果。MoCo[9]采用动量编码器机制,将一批负样本的特征表示存储在队列中,并使用动量更新来维持表示的一致性。这使模型能够有效地保留不同小批量之间的语义信息,解决了传统队列对比学习对大批量负样本的依赖问题。与其他对比学习方法不同,BYOL[4]不需要显式的负样本。它通过 2 个网络进行训练:在线网络和目标网络。在线网络通过梯度下降更新,而目标网络通过指数移动平均方法更新。通过将增强版本的视图投射到目标空间,并与在线网络生成的另一视图对齐来学习高质量的表征。其创新之处在于消除了对负样本的需求,同时获得了较好的结果。MoCo-v3[10]是 MoCo 系列的改进版本,基于 MoCo-v2[11]进行了重要更新。它整合了 Transformer 架构,并通过改进的数据增强和优化策略,在自监督视觉学习任务中实现了更稳定、更高效的性能。

除此之外,基于掩码技术的自监督学习方法在多个领域得到广泛应用。通过将输入数据中的一部分信息隐藏或遮盖,掩码技术迫使模型通过上下文和周围信息来填补缺失部分。这一方法有效模拟了无标签数据的学习情境,使得模型能够在缺乏明确监督信号的情况下,从数据中自动提取有用特征。在 BERT[12]中,研究者通过掩盖文本中的部分单词,引导模型根据上下文语义预测这些被掩蔽的内容。而在 MAE[13]中,模型通过对随机遮蔽的图像区域进行重建,从而学习到更为深层次、语义丰富的视觉特征表示。MFM[14]则提出了一种创新性的自监督预训练方法,即掩码频率建模。与传统的空间域掩码方法不同,它将掩蔽操作应用于图像的频率成分,并训练模型预测这些缺失的频率信息。实验结果表明,掩码频率建模在性能稳定性和抗干扰能力方面均优于传统的掩码图像建模技术,并在图像分类和语义分割等多种视觉任务中展现出卓越性能。

然而,上述这些方法在艺术图像领域的适用性尚未得到充分验证。首先,现有的自监督学习方法主要针对自然图像设计,未充分考虑艺术图像的独特性。艺术图像相比自然图像具有更为复杂的纹理结构、风格多样性以及抽象表达方式,使得常规的数据增强和特征提取策略难以有效捕捉艺术作品的本质特征。其次,现有的数据增强方法主要集中在颜色变换、几何变形等低层次视觉特征上,而艺术图像的风格、构图和意境等高层次语义信息往往被忽视。这导致模型在学习艺术图像表征时,过度关注表面特征而忽略了艺术作品的深层语义内涵,限制了分类性能的进一步提升。最后,艺术图像数据集通常规模有限且类别分布不均衡,使得基于大规模数据训练的自监督学习方法难以直接迁移应用。如何在有限数据条件下实现高效的特征学习,成为艺术图像分类中的关键挑战。

针对上述问题,提出一种基于小波变换增强的双

分支对比学习方法,其重要意义体现在以下几个方面:

① 理论创新方面。创新性地将小波变换作为一种特殊的数据增强手段引入自监督学习框架,为艺术图像特征提取提供了新的视角。小波变换能够有效捕捉图像的多尺度特征和纹理信息,这与艺术图像的本质特性高度契合,有助于模型学习更加全面且判别性强的艺术图像表征。

② 技术方法方面。采用的双分支对比学习架构,巧妙地结合了常规数据增强和小波变换增强2种不同的特征提取路径,使模型能够同时学习艺术图像的不同层次表征。这种架构设计不仅提高了模型的特征提取能力,还增强了其对艺术风格和内容的理解深度。

③ 应用价值方面。所提方法在艺术图像分类任务上的性能提升,为数字艺术研究、文化遗产保护和艺术教育等领域提供了有力的技术支持。尤其在样本数量有限的情况下,该方法通过自监督学习有效减少了对标注数据的依赖,具有显著的实用价值。

此外,所提出的框架具有良好的泛化性和扩展性,不仅适用于艺术图像分类,还可推广到其他特殊领域的视觉任务中,为计算机视觉在专业领域的应用提供了新的思路和解决方案。

1 整体框架及相关技术介绍

1.1 整体框架

F-MaCo整体框架如图1所示,它是一种基于动量更新的双分支对比学习结构,并结合频域掩码增强和感知损失权重机制,提升对艺术图像细节及其结构的建模能力。

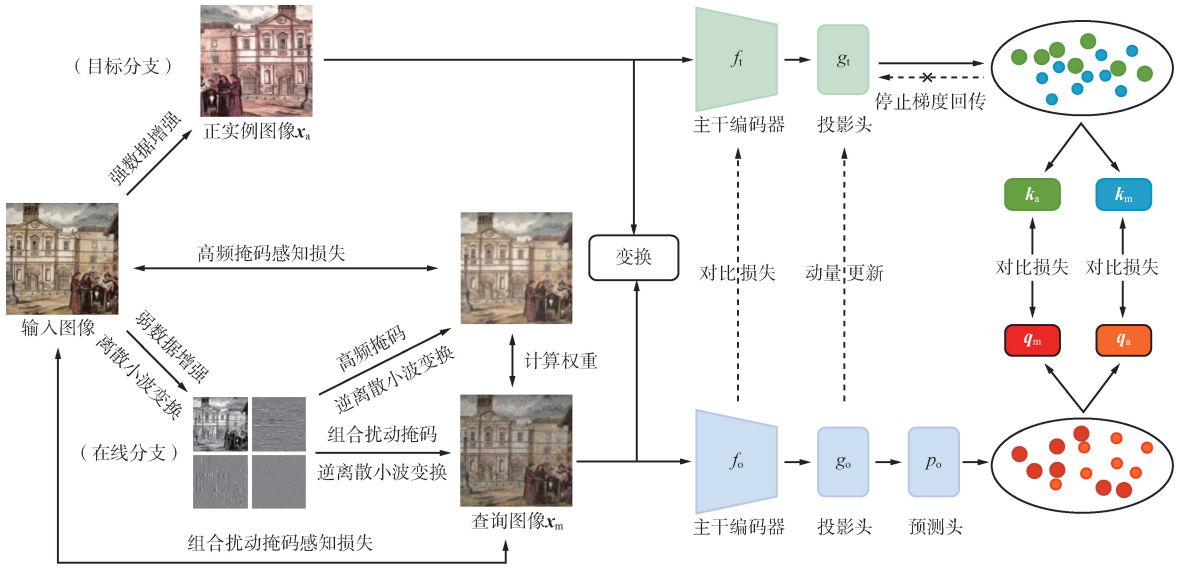


图1 F-MaCo整体框架

Fig. 1 Overall framework of F-MaCo

1.1.1 双分支架构

模型包括在线分支与目标分支。在线分支包含基于Vision Transformer(ViT)架构的主干编码器f_o、投影头g_o以及预测头p_o,用于生成查询特征$\boldsymbol{q}$。目标分支包含与在线分支结构相同但参数不同的主干编码器f_t和投影头g_t,用于生成对比特征$\boldsymbol{k}$。目标分支的参数通过在线分支以动量更新的方式得到:

$$\theta_t = m \cdot \theta_t + (1 - m) \cdot \theta_o, \tag{1}$$

式中:θ_t和θ_o分别表示目标分支和在线分支的参数,$m \in [0,1)$为动量系数。

1.1.2 频域掩码与强数据增强

一方面,输入图像经过DWT后,施加2种频域扰动掩码策略,包括高频子带采样掩码和组合扰动掩码,得到高频掩码图像和组合扰动掩码图像,其中高频掩码图像仅用于计算感知损失,组合扰动掩码图像输入到在线分支中;另一方面,原图经过强数据增强后得到的增强视图输入到目标分支中。

1.1.3 主干编码与对比计算

掩码视图$\boldsymbol{x}_m$输入到在线分支的主干编码器,提取其语义特征表示,随后经过投影头和预测头生成查询特征$\boldsymbol{q}_m$,增强视图$\boldsymbol{x}_a$则输入到目标分支的编码器,经过投影头得到对应样本特征$\boldsymbol{k}_a$。随后交换2条分支的输入,将掩码视图输入到目标分支中得

到特征 $\boldsymbol{k}_{\mathrm{m}}$;将增强视图输入到在线分支中得到特征 $\boldsymbol{q}_{\mathrm{a}}$。在此基础上分别构造 2 个对比项:以 $\boldsymbol{q}_{\mathrm{m}}$ 为查询,$\boldsymbol{k}_{\mathrm{a}}$ 为正样本构造标准对比项;以 $\boldsymbol{q}_{\mathrm{a}}$ 为查询,$\boldsymbol{k}_{\mathrm{m}}$ 为正样本构造交换视图下的对比项。这种互换结构可以进一步鼓励模型在多种扰动条件下保持语义对齐,提升表征的稳定性与判别力。

1.1.4 动态权重计算与加权对比损失

通过引入感知损失衡量不同掩码重建图像的结构保真度差异,计算动态权重。对于每一对由在线分支生成的查询特征与目标分支生成的对应样本特征构成的查询,即键对,计算其与所有负样本之间的对比损失,并引入感知误差计算出动态权重,最终损失定义为加权后的双分支对比损失之和。

1.2 ViT

随着自然语言处理领域中 Transformer 架构的成功应用,ViT[15] 被提出作为一种处理图像任务的新型深度学习架构。相比于传统的卷积神经网络,ViT 通过自注意力机制直接对图像进行全局建模,在图像分类等任务中展现出卓越的性能。

Transformer 架构示意如图 2 所示,ViT 的核心思想是将图像视为一系列的图像块序列,并应用 Transformer 架构处理这些序列。首先将输入图像 $\boldsymbol{x}\in\mathbb{R}^{H\times W\times C}$ 分割成固定大小的图像块,H 和 W 分别表示图像的高度和宽度,C 为通道数,然后通过线性投影映射到特征空间:

$$\boldsymbol{z}_0=[\boldsymbol{x}_{\mathrm{cls}};\boldsymbol{x}_{\mathrm{p}}^1\boldsymbol{E};\boldsymbol{x}_{\mathrm{p}}^2\boldsymbol{E};\cdots;\boldsymbol{x}_{\mathrm{p}}^N\boldsymbol{E}]+\boldsymbol{E}_{\mathrm{pos}},\tag{2}$$

式中:$\boldsymbol{x}_{\mathrm{p}}$ 为图像块序列,$\boldsymbol{x}_{\mathrm{p}}^i$ 表示第 i 个图像块,共有 N 个图像块;$\boldsymbol{E}$ 为将图像块映射到特征空间的线性投影矩阵,$\boldsymbol{x}_{\mathrm{cls}}$ 为可学习的分类标记,用于最终的分类任务;$\boldsymbol{E}_{\mathrm{pos}}$ 为位置编码,用于保留图像块的位置信息。ViT 中包含 L 个相同的层,每层由多头自注意力机制和多层感知机组成:

$$\boldsymbol{z}_l'=MSA(LN(\boldsymbol{z}_{l-1}))+\boldsymbol{z}_{l-1},\tag{3}$$

$$\boldsymbol{z}_l=MLP(LN(\boldsymbol{z}_l'))+\boldsymbol{z}_l',\tag{4}$$

式中:$\boldsymbol{z}_l$ 表示第 l 层的输出特征,MSA 表示多头自注意力机制,LN 表示层归一化操作,MLP 表示多层感知机。ViT 的多头自注意力机制允许模型捕捉输入序列中相距较远的元素之间存在的重要关联:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V},\tag{5}$$

式中:$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别表示查询、键和值矩阵,d_k 为键向量的维度,用于缩放点积注意力。

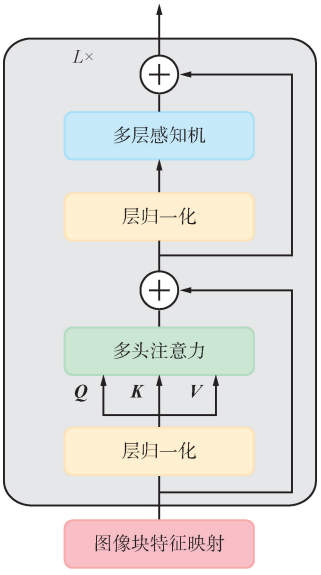


图 2 Transformer 架构示意

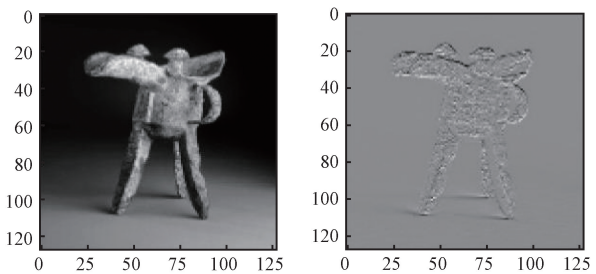
Fig. 2 Schematic of Transformer architecture

1.3 DWT 和二维逆离散小波变换

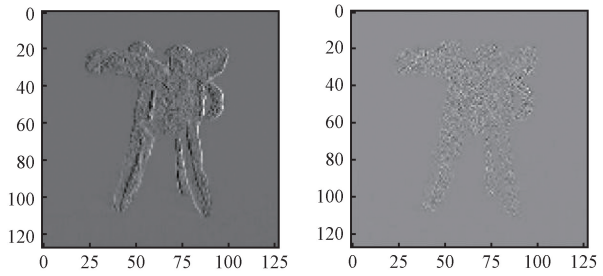
傅里叶变换在频域上具有良好的全局频率分析能力,但缺乏对空间局部特征的刻画能力,难以有效捕捉图像中局部的纹理与边缘信息。而小波变换具备良好的时频局部化特性,能够在多尺度上同时分析图像的整体构图与细节纹理,特别适用于反映艺术图像中丰富的笔触变化、材质特征及风格细节。因此,为了提升模型对艺术图像中复杂纹理、边缘结构及风格特征的提取能力,采用 DWT 将图像从空间域转换到频域表示,便于分别掩码其高频与低频信息。在掩码处理后,通过二维逆离散小波变换(Inverse Discrete Wavelet Transform,IDWT)将其还原回原始空间域。

1.3.1 DWT

小波变换各频带的灰度示意如图 3 所示,DWT 通过在水平方向与垂直方向上分别施加一维小波变换,将输入图像分解为 4 个频率子带:低频分量($\boldsymbol{LL}$)与 3 个方向的高频分量($\boldsymbol{LH}$、$\boldsymbol{HL}$、$\boldsymbol{HH}$)。



(a) 低频子带　　(b) 水平高频子带



(c) 垂直高频子带　　(d) 对角高频子带

图 3　小波变换各频带的灰度示意

Fig. 3　Gray-scale representation of wavelet transform subbands

对每个通道 $\boldsymbol{x}_c \in \mathbb{R}^{H\times W}$ 单独进行小波分解,设 $\boldsymbol{h}_{\mathrm{L}}$ 和 $\boldsymbol{h}_{\mathrm{H}}$ 分别为低通和高通滤波器,DWT 的变换过程可表示为:

$$\boldsymbol{LL} = (\boldsymbol{x}_c * \boldsymbol{h}_{\mathrm{L}}^{\mathrm{T}}\boldsymbol{h}_{\mathrm{L}}) \downarrow 2, \tag{6}$$

$$\boldsymbol{LH} = (\boldsymbol{x}_c * \boldsymbol{h}_{\mathrm{L}}^{\mathrm{T}}\boldsymbol{h}_{\mathrm{H}}) \downarrow 2, \tag{7}$$

$$\boldsymbol{HL} = (\boldsymbol{x}_c * \boldsymbol{h}_{\mathrm{H}}^{\mathrm{T}}\boldsymbol{h}_{\mathrm{L}}) \downarrow 2, \tag{8}$$

$$\boldsymbol{HH} = (\boldsymbol{x}_c * \boldsymbol{h}_{\mathrm{H}}^{\mathrm{T}}\boldsymbol{h}_{\mathrm{H}}) \downarrow 2, \tag{9}$$

式中:$*$ 表示二维卷积操作,$\downarrow 2$ 表示对行列方向分别进行 2 倍下采样,$\boldsymbol{LL} \in \mathbb{R}^{H/2\times W/2}$ 表示图像的低频分量,保留图像整体结构,$\boldsymbol{LH}$ 表示水平方向的高频边缘,$\boldsymbol{HL}$ 表示垂直方向的高频边缘,$\boldsymbol{HH}$ 表示对角线方向的细节纹理。

1.3.2　IDWT

IDWT 可将上述 4 个频带图像重新合成为原始图像。其过程包括对每个子带进行上采样,再与对应的重构滤波器进行卷积,求和得到重构图像。具体公式为:

$$\hat{\boldsymbol{x}}_c = (\uparrow 2\ \boldsymbol{LL} * \boldsymbol{h}_{\mathrm{L}}^{\mathrm{T}}\boldsymbol{h}_{\mathrm{L}}) + (\uparrow 2\ \boldsymbol{LH} * \boldsymbol{h}_{\mathrm{L}}^{\mathrm{T}}\boldsymbol{h}_{\mathrm{H}}) + (\uparrow 2\ \boldsymbol{HL} * \boldsymbol{h}_{\mathrm{H}}^{\mathrm{T}}\boldsymbol{h}_{\mathrm{L}}) + (\uparrow 2\ \boldsymbol{HH} * \boldsymbol{h}_{\mathrm{H}}^{\mathrm{T}}\boldsymbol{h}_{\mathrm{H}}), \tag{10}$$

式中:$\uparrow 2\ X$ 表示将分量 X 进行 2 倍上采样,$\hat{\boldsymbol{x}}_c \in \mathbb{R}^{H\times W}$ 为第 c 通道重构后的图像,重构结果 $\hat{\boldsymbol{x}} \in \mathbb{R}^{H\times W}$ 即为完整的图像恢复版本。

1.4　动态频域掩码

为增强模型对频率信息变化的鲁棒性,在图像经小波变换后的频域表示中引入动态频域掩码机制,用于生成多样化的输入样本,提升模型对不同频率特征的辨识能力。高频掩码与组合扰动掩码如图 4 所示,提出一种随训练阶段动态调整扰动强度的组合扰动掩码机制,对低频与高频分量分别采用不同的扰动方式,增强模型在训练过程中的稳健性与泛化能力。

(a) 原图　　(b) 高频掩码　　(c) 组合扰动掩码

图 4　高频掩码与组合扰动掩码

Fig. 4　High-frequency mask and combined perturbation mask

对输入图像的第 c 个通道施加 DWT,可得:

$$DWT(\boldsymbol{x}_c) = \{\boldsymbol{LL}_c, \boldsymbol{LH}_c, \boldsymbol{HL}_c, \boldsymbol{HH}_c\}。 \tag{11}$$

掩码形式为:

$$\mathcal{M}_{\mathrm{random}}(\boldsymbol{x}_c^{(t)}) = \{\widetilde{\boldsymbol{LL}}_c^{(t)}, \widetilde{\boldsymbol{LH}}_c^{(t)}, \widetilde{\boldsymbol{HL}}_c^{(t)}, \widetilde{\boldsymbol{HH}}_c^{(t)}\}, \tag{12}$$

式中:t 表示当前训练轮次。高频掩码通过如下步骤实施:对高频 3 个子带进行统一采样,形成一个总的索引掩码,再调整其形状,分别作用到 3 个分量中。

$$\boldsymbol{m}^{(t)} = Sample(1 - r^{(t)}),\ \boldsymbol{m}^{(t)} \in \{0,1\}^{|\boldsymbol{LH}|+|\boldsymbol{HL}|+|\boldsymbol{HH}|}, \tag{13}$$

$$\widetilde{\boldsymbol{LH}}_c^{(t)} = \boldsymbol{LH}_c \odot \boldsymbol{m}_{\boldsymbol{LH}}^{(t)}, \tag{14}$$

$$\widetilde{\boldsymbol{HL}}_c^{(t)} = \boldsymbol{HL}_c \odot \boldsymbol{m}_{\boldsymbol{HL}}^{(t)}, \tag{15}$$

$$\widetilde{\boldsymbol{HH}}_c^{(t)} = \boldsymbol{HH}_c \odot \boldsymbol{m}_{\boldsymbol{HH}}^{(t)}, \tag{16}$$

式中:$Sample$ 表示随机采样,$\odot$ 表示逐元素乘法,$\boldsymbol{m}$ 为包含了 0 和 1 的掩码向量。掩码率采用余弦调度函数:

$$r^{(t)} = r_{\mathrm{start}} + (r_{\mathrm{end}} - r_{\mathrm{start}}) \cdot \frac{1 - \cos\left(\pi \cdot \frac{t}{T}\right)}{2}, \tag{17}$$

式中:$r^{(t)}$ 表示第 t 轮的高频掩码率,r_{start} 和 r_{end} 表示初始与最终的掩码率,T 表示总训练轮数。这种策略使得在训练初期信息保留较多,有利于模型稳定学习,而训练后期扰动增强,有助于泛化能力提升。在低频分量中,采用 3 个线性变化的扰动参数来控制软掩码强度、缩放幅度与噪声强度:

$$\mu^{(t)} = 0.9 - 0.3 \cdot \frac{t}{T}, \mathcal{M}_{\mathrm{soft}}^{(t)} \sim \mathcal{U}(\mu^{(t)}, 1), \tag{18}$$

$$\alpha^{(t)} = 1.0 - 0.2 \cdot \frac{t}{T}, \tag{19}$$

$$\sigma^{(t)} = 0.005 + 0.015 \cdot \frac{t}{T}, \tag{20}$$

式中:$\mu^{(t)}$ 为第 t 轮的软掩码强度下限,$\mathcal{M}_{\mathrm{soft}}^{(t)}$ 为第

t 轮的软掩码强度,$\mathcal{U}(\cdot)$ 为均匀采样,$\alpha^{(t)}$ 为第 t 轮的缩放幅度,$\sigma^{(t)}$ 第 t 轮的噪声标准差。最终扰动后的低频项为:

$$\widetilde{\boldsymbol{LL}}_c^{(t)} = \alpha^{(t)} \cdot (\boldsymbol{LL}_c \odot \boldsymbol{\mathcal{M}}_{\text{soft}}^{(t)}) + \boldsymbol{\epsilon}^{(t)},\ \boldsymbol{\epsilon}^{(t)} \sim \mathcal{N}(0,(\sigma^{(t)})^2), \tag{21}$$

式中:$\mathcal{N}(\cdot)$ 为正态分布。为避免扰动过强导致图像结构失真或训练过程中的不稳定性,软掩码的扰动强度、缩放系数以及加性噪声的标准差,均在预设范围内随训练进程缓慢变化,起始阶段保持较低扰动水平,以保留图像主干信息并促进稳定特征提取。训练过程中逐步加强扰动强度,调节频域信息保留程度,引导模型从全局结构到局部细节逐步建立鲁棒的表示能力,为后续感知损失与表征对齐机制奠定基础。

1.5 权重调整

在前述频域掩码增强基础上,进一步引入感知损失驱动的动态权重调整机制,用于提升不同扰动策略在训练过程中的相对贡献性。该机制以感知误差为反馈信号,自动调节损失中掩码图像与原图之间的学习比重,从而实现针对性更强的增强学习。采用基于 VGG16 网络的感知损失衡量图像结构差异,定义如下:

$$\mathcal{L}_{\text{perc}}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \|\phi(\boldsymbol{x}) - \phi(\hat{\boldsymbol{x}})\|_2^2, \tag{22}$$

式中:$\phi(\cdot)$ 表示提取的 VGG 中间层特征,$\boldsymbol{x}$ 表示原始图像,$\hat{\boldsymbol{x}}$ 表示掩码扰动后的图像,感知损失按样本维度平均,保留每个样本的误差大小。计算组合扰动掩码和高频掩码的感知损失,并设计动态权重函数,具体如下:高频掩码通过抑制高频分量,仅保留低频子带 $\boldsymbol{LL}_c$,从而构造一幅仅包含主干结构信息的图像。其频域掩码定义为:

$$\boldsymbol{\mathcal{M}}_{\text{high}}(\boldsymbol{x}) = \{\boldsymbol{LL},0,0,0\}。 \tag{23}$$

对应还原图像为:

$$\tilde{\boldsymbol{x}}_{\text{high}} = IDWT(\boldsymbol{\mathcal{M}}_{\text{high}}(\boldsymbol{x}))。 \tag{24}$$

组合扰动掩码对应还原图像为:

$$\tilde{\boldsymbol{x}}_{\text{random}} = IDWT(\boldsymbol{\mathcal{M}}_{\text{random}}(\boldsymbol{x}))。 \tag{25}$$

分别计算二者与原图之间的感知损失:

$$e_{\text{high}} = \mathcal{L}_{\text{perc}}(\boldsymbol{x},\tilde{\boldsymbol{x}}_{\text{high}}), \tag{26}$$

$$e_{\text{random}} = \mathcal{L}_{\text{perc}}(\boldsymbol{x},\tilde{\boldsymbol{x}}_{\text{random}})。 \tag{27}$$

为合理评估组合扰动掩码的训练贡献,设计动态权重函数:

$$\omega = clip\left(\frac{e_{\text{random}}}{e_{\text{high}}+\varepsilon},\omega_{\min},\omega_{\max}\right), \tag{28}$$

式中:$\omega_{\min}=0.5$、$\omega_{\max}=2.0$ 控制权重上下限,避免训练不稳定,$\varepsilon=10^{-6}$ 为数值稳定性常数,最终将 ω 应用于后续损失函数中对应样本的权重计算。

该机制可视为一种基于误差的难度引导策略:若组合掩码后的图像更难还原(感知误差大),则给予其更高的权重,鼓励模型学习具有挑战性的特征结构;反之,则减弱其影响。之所以选取高频子带图像作为感知误差的计算基础,主要考虑其在图像结构中的独特作用。高频部分集中反映图像中的边缘、纹理和细节变化,这些元素在艺术图像的风格体现与形式差异中具有重要意义。相较于低频信息所反映的轮廓与构图,高频区域对扰动更加敏感,其差异更能揭示结构保真度的变化。因此,以高频子带作为感知差异的度量对象,有助于更准确地反映掩码策略对图像结构的干扰程度,并据此调节对比学习过程中的样本权重,提高对关键纹理信息的学习效率。

1.6 损失函数设计

基于 InfoNCE[8-9, 16] 损失,引入动态样本权重机制,以提升不同频域扰动样本在训练中的有效性。给定一对正样本特征 $(\boldsymbol{q},\boldsymbol{k}^+)$,以及来自当前批次的其他负样本特征集合 $\{\boldsymbol{k}^-\}$,传统的 InfoNCE 损失为:

$$\mathcal{L}_{\text{InfoNCE}} = -\lg\frac{\exp(\boldsymbol{q}\cdot\boldsymbol{k}^+/\tau)}{\sum_{\boldsymbol{k}^-}\exp(\boldsymbol{q}\cdot\boldsymbol{k}^-/\tau)}, \tag{29}$$

式中:τ 为温度系数。对每个正样本分配一个动态权重 ω,并在损失函数外引入额外的全局缩放系数 $2\cdot\tau$。提升与感知损失的数值尺度对齐能力,从而定义实际使用的损失:

$$\mathcal{L}_{q,k} = -2\cdot\tau\cdot\omega\cdot\lg\frac{\exp\left(\boldsymbol{q}\cdot\frac{\boldsymbol{k}^+}{\tau}\right)}{\sum_{\boldsymbol{k}^-}\exp\left(\boldsymbol{q}\cdot\frac{\boldsymbol{k}^-}{\tau}\right)}。 \tag{30}$$

以双分支对比学习方式进行训练,计算交叉损失:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{q_1,k_2} + \mathcal{L}_{q_2,k_1}。 \tag{31}$$

2 实验设计与结果分析

2.1 硬件配置

所有算法的训练与评估均在同一硬件平台上进行,以保证实验结果的公平性和可比性。具体而言,使用了 3 块 NVIDIA GeForce RTX 3090 GPU 进行所有实验,该显卡具备强大的计算能力和充足的显存资源(每块 24 GB),能够支持大规模图像输入与高效的训练过程。

2.2 实验数据集介绍

本研究在4个常用的艺术图像分类基准数据集上进行了系统性的实验验证,涵盖了从小型到大型、从低难度到高难度的多种任务场景,旨在全面评估所提方法在不同条件下的分类性能与跨数据集的泛化能力。图5展示了各数据集中部分具有代表性的图像示例。其中,MAMe[17]是一个专注于艺术品媒介识别的高分辨率和可变形状的数据集,图像内容丰富,含有37 407张图片,分成29个类别;Kaokore[18]是一个前现代日本艺术面部表情数据集,含有7 294张图片,分成4个类别,主要挑战在于表情间的细微差别及图像风格的一致性;Artbench10[19]包含60 000张艺术作品图像,涵盖了10种不同的艺术风格,是当前艺术风格分类领域广泛使用的标准数据集;ArtDL[20]是一个用于图像分类的新型绘画数据集,含有26 413张图片,分成19个类别覆盖多种题材和风格,适合用于模型的泛化能力测试。这些数据集的详细统计信息如表1所示,作为后续实验分析与对比的基础。

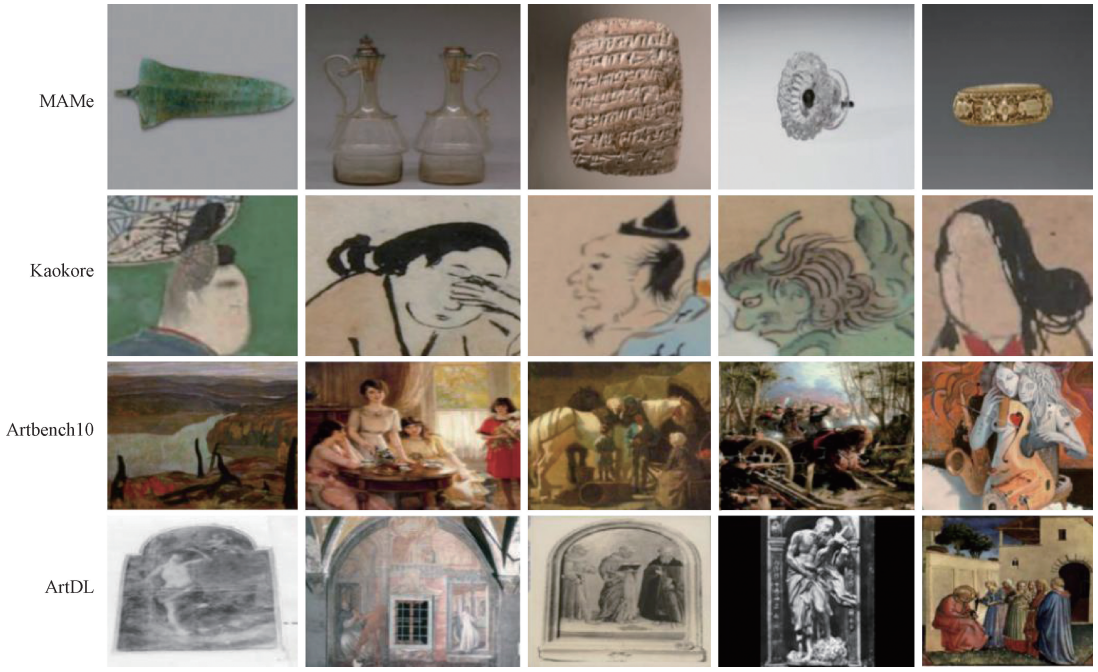


图5 数据集中的部分示例

Fig. 5 Partial examples in the datasets

表1 数据集详情

Tab. 1 Dataset details

数据集	类别	图片尺寸/pixel	训练集数量	测试集数量	验证集数量
MAMe	29	256×256	20 300	15 657	1 450
Kaokore	4	256×256	5 815	745	734
Artbench10	10	256×256	50 000	10 000	—
ArtDL	19	不唯一	20 619	2 897	2 897

2.3 实验参数设置

采用ViT-S[10]作为骨干网络、AdamW[21]优化器进行训练。在训练过程中,设置初始高频掩码率 $r_{start}=25\%$,最终高频掩码率 $r_{end}=75\%$ 。此外,引入了余弦退火学习率调度,并在前40轮采用学习率热身策略,避免模型在早期更新过快导致的不稳定性。详细参数设置如表2所示。

表2 参数设置

Tab. 2 Parameter settings

参数	值	参数	值
学习率	0.000 15	训练总轮数	300
权重衰减	0.1	分类器	SGD
批次大小	256	是否包含BN层	否
动量	0.9	分类器学习率	0.1

续表

参数	值	参数	值
优化器	AdamW	分类器权重衰减	0
优化器动量因子	(0.9,0.95)	评估轮数	90
热身轮数	40		

在模型性能评估阶段,采用自监督学习中常用的线性探测和 K 近邻分类 2 种方法进行模型评估,评估指标均为 Top-1 准确率。Top-1 准确率表示预测概率最高的类别与真实类别一致的比例。

2.4 实验结果及分析

对所提出方法在多个数据集上进行了评估,包括线性探测和 K 近邻分类两类实验,以全面衡量其在图像表征学习方面的效果。对于存在官方训练集与验证集划分的数据集,分别使用训练集进行特征学习,使用验证集进行评估。对于未提供验证集划分的数据集,在测试集上进行评估。

2.4.1 线性探测

为确保不同方法之间评估条件的一致性,在进行线性探测实验时,采用了最简化的分类器设计:仅在预训练网络的输出特征之后添加一个全连接层,且不包含 Batch Normalization、激活函数或其他非线性处理操作。这样的设置可直接衡量图像特征在原始空间中的可分性,避免评估结果受到额外结构的影响。

该实验策略在无监督图像表征研究中已被广泛采用,是目前衡量特征分离能力的常规做法。实验结果如表 3 所示。可以看出,F-MaCo 方法在所有 4 个数据集上均取得了最优的性能表现,在 MAMe 数据集上,F-MaCo 取得了 73.72%的准确率,相比次优精度提升了 0.69%。在 Kaokore 数据集上,F-MaCo 准确率为 77.38%,相比次优精度提升了 1.08%。在 Artbench10 数据集上,F-MaCo 达到 58.38%,相比次优精度提升了 0.72%。在 ArtDL 数据集上,F-MaCo 准确率为 68.31%,相比次优精度提升了 0.24%。

表 3 不同自监督学习方法在 4 个数据集上的 Top-1 准确率结果

Tab. 3 Top-1 accuracy results of different self-supervised learning methods on four datasets

单位:%

方法	轮数	模型	MAMe *Top*-1	Kaokore *Top*-1	Artbench10 *Top*-1	ArtDL *Top*-1
BEiT[22]	300	ViT-S	71.31	73.98	51.37	66.45
MAE[13]	300	ViT-S	55.72	63.08	45.36	65.28
CAE[23]	300	ViT-S	72.73	75.88	53.95	67.93
BYOL[4]	300	ViT-S	68.97	72.07	54.30	66.48
SimSiam[24]	300	ViT-S	64.07	71.39	46.62	66.41
MoCo-v3[10]	300	ViT-S	73.03	76.30	57.66	68.07
DINO[25]	300	ViT-S	72.00	66.62	53.80	66.80
F-MaCo	300	ViT-S	**73.72**	**77.38**	**58.38**	**68.31**

2.4.2 K 近邻分类

线性探测结果清楚地表明,F-MaCo 明显优于大多数经典方法。为了进一步验证所学特征在无监督评估条件下的泛化能力与稳定性,引入 K 近邻分类实验进行辅助评估,其中 K 分别设为 20 与 100。该方法无需额外训练分类器,直接基于预训练特征进行多数投票预测,能够有效衡量模型在不依赖下游监督信号时的表征鲁棒性。如表 4 所示,在 K 近邻分类实验中,F-MaCo 的表现比最先进的方法高出 0.34%~3.55%,表明其预训练阶段学到的特征具备更强的可迁移性与表达能力,能够在无需微调的条件下实现更优的下游分类性能。

表 4 不同自监督学习方法在 4 个数据集上的 KNN 分类性能

Tab. 4 KNN classification performance of different self-supervised learning methods on four datasets

单位:%

方法	MAMe		Kaokore		Artbench10		ArtDL	
	$K=20$	$K=100$	$K=20$	$K=100$	$K=20$	$K=100$	$K=20$	$K=100$
BEiT[22]	60.34	55.24	70.03	67.30	39.61	39.28	66.31	65.28
MAE[13]	45.45	40.28	59.95	58.45	40.84	40.00	66.21	66.28
CAE[23]	65.52	59.45	75.75	72.34	49.28	48.07	66.69	66.45
BYOL[4]	69.52	65.38	75.61	72.21	52.84	51.41	66.72	66.34

续表

方法	MAMe		Kaokore		Artbench10		ArtDL	
	$K=20$	$K=100$	$K=20$	$K=100$	$K=20$	$K=100$	$K=20$	$K=100$
SimSiam[24]	66.48	60.62	77.93	72.89	46.28	44.97	66.21	65.48
MoCo-v3[10]	73.17	70.28	79.15	78.75	54.37	54.82	67.59	67.69
DINO[25]	73.17	71.24	66.49	65.80	53.57	53.36	67.59	66.93
F-Maco	**75.24**	**74.00**	**82.70**	**80.38**	**57.50**	**57.00**	**68.86**	**68.03**

综上可知,所提出的 F-MaCo 方法在艺术图像表征学习任务中的有效性和鲁棒性明显优于其他主流自监督学习方法。

2.5 消融实验

为验证频域掩码对 F-MaCo 整体性能的贡献,在 Kaokore 数据集上进行了消融实验,将频域掩码分别应用在在线分支和目标分支中,并在相同的实验参数设置下进行训练与测试。如表 5 所示,频域掩码对模型性能具有显著影响,尤其在在线分支中的引入效果更优,表明该机制在增强模型判别能力方面发挥了关键作用。此外,目标分支引入频域掩码虽然在性能上提升较小,但仍高于不使用频域信息的情况,验证了频域掩码的有效性与可扩展性。

表 5 比较频域掩码对在线分支与目标分支的影响

Tab. 5 Comparison of the effects of frequency domain masking on the online and target branches

单位:%

在线分支		目标分支		*Top*-1
强增强	弱+频域掩码	强增强	弱+频域掩码	
×	×	×	×	51.92
√	×	√	×	76.30
×	√	√	×	77.38
√	×	×	√	77.00

为进一步探索动态频域掩码策略对模型性能的影响,设计了不同的起始掩码率 r_{start} 和最终掩码率 r_{end} 的组合方案,并观察其对最终 Top-1 准确率的影响。不同动态掩码率设置对模型性能的影响如表 6 所示,当掩码率从 $r_{start}=0$ 动态增长至 $r_{end}=0.25$ 和 $r_{end}=0.50$ 时,模型性能分别达到 77.05%和 77.14%,性能略有提升;当起始掩码率提高至 $r_{start}=0.25$ 并最终增长至 $r_{end}=0.75$ 时,模型准确率进一步提升至 77.38%。上述结果表明,适当提高初始掩码强度并保持一定的增长幅度,有助于模型在训练过程中逐步适应更复杂的频域干扰,从而获得更稳健的特征表征能力。这一动态策略相比静态频域掩码更具灵活性,能够在保持信息多样性的同时,有效增强模型的判别性能。

表 6 不同动态掩码率设置对模型性能的影响

Tab. 6 Efects of different dynamic masking rates on model performance

r_{start}	r_{end}	*Top*-1/%
0	0.25	77.05
0	0.50	77.14
0.25	0.75	77.38

3 结束语

针对艺术图像分类提出了一种基于小波变换增强的双分支对比学习方法,通过创新融合频域分析与对比学习机制,为艺术图像表征学习开辟了新路径。本研究突破传统自监督学习依赖自然图像设计范式的局限,利用 DWT 挖掘艺术图像多尺度纹理与结构特征,并通过动态频域掩码机制在训练过程中渐进式增强模型对频率特征的鲁棒性,该策略使模型在信息保留与扰动增强间实现动态平衡,有效提升了对艺术图像复杂风格与抽象表达的特征捕捉能力。实验结果表明,该方法在多种类型艺术图像数据集上实现了性能突破,验证了其在特征学习与表征泛化能力上的显著优势。

理论价值在于为艺术图像自监督学习提供了频域增强的新思路,突破了传统空间域数据增强的局限性,不仅为数字艺术研究、文化遗产保护等领域提供了实用技术支撑,更重要的是为特殊领域视觉任务的自监督学习提供了可迁移的方法论框架,推动计算机视觉在专业领域的研究向更深层次发展。

参 考 文 献

[1] CARNEIRO G, SILVA N P D, BUE A D, et al. Artistic Image Classification: An Analysis on the PRINTART Database[C]//European Conference on Computer Vision. Florence: Springer, 2012: 143-157.

[2] RODRIGUEZ C S, LECH M, PIROGOVA E. Classification of Style in Fine-art Paintings Using Transfer Learning and Weighted Image Patches[C]//International Conference on Signal Processing and Communication Systems. Queensland: IEEE, 2018: 1-7.

[3] OZBULAK U,LEE H J,BOGA B,et al. Know Your Self-supervised Learning:A Survey on Image-based Generative and Discriminative Training[EB/OL]. (2023-05-23)[2025-05-18]. https://arxiv.org/abs/2305.13689.

[4] GRILL J B,STRUB F,ALTCHE F,et al. Bootstrap Your Own Latent a New Approach to Self-supervised Learning[C]//Advances in Neural Information Processing Systems. Vancouver:Curran Associates Inc.,2020:21271-21284.

[5] ZBONTAR J,JING L,MISRA I,et al. Barlow Twins:Self-supervised Learning via Redundancy Reduction[C]//International Conference on Machine Learning. Vienna:IMLS,2021:12310-12320.

[6] CARON M,MISRA I,MAIRAL J,et al. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments[C]//Advances in Neural Information Processing Systems. Vancouver:Curran Associates Inc.,2020:9912-9924.

[7] WU Z R,XIONG Y J,YU S X,et al. Unsupervised Feature Learning via Non-parametric Instance Discrimination[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE,2018:3733-3742.

[8] CHEN T,KORNBLITH S,NOROUZI M,et al. A Simple Framework for Contrastive Learning of Visual Representations[C]//International Conference on Machine Learning. Online:JMLR,2020:1597-1607.

[9] HE K M,FAN H Q,WU Y,et al. Momentum Contrast for Unsupervised Visual Representation Learning[C]//IEEE Conference on Computer Vision and Pattern Recognition. Seattle:IEEE,2020:9729-9738.

[10] CHEN X L,XIE S N,HE K M. An Empirical Study of Training Self-supervised Vision Transformers[C]//IEEE International Conference on Computer Vision. Montreal:IEEE,2021:9640-9649.

[11] CHEN X L,FAN H Q,GIRSHICK R,et al. Improved Baselines with Momentum Contrastive Learning[EB/OL]. (2020-03-09)[2025-05-18]. https://arxiv.org/abs/2003.0429.

[12] DEVLIN J,CHANG M W,LEE K,et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//The Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. Minneapolis:ACL,2019:4171-4186.

[13] HE K M,CHEN X L,XIE S N,et al. Masked Autoencoders Are Scalable Vision Learners[C]//IEEE Conference on Computer Vision and Pattern Recognition. New Orleans:IEEE,2022:16000-16009.

[14] XIE J H,LI W,ZHAN X H,et al. Masked Frequency Modeling for Self-supervised Visual Pre-training[EB/OL]. (2023-04-25)[2025-05-18]. https://arxiv.org/abs/2206.07706.

[15] DOSOVITSKIY A,BEYER L,KOLESNIKOV A,et al. An Image Is Worth 16×16 Words:Transformers for Image Recognition at Scale[EB/OL]. (2021-06-03)[2025-05-18]. https://arxiv.org/abs/2010.11929.

[16] OORD A,LI Y,VINYALS O. Representation Learning with Contrastive Predictive Coding[EB/OL]. (2019-01-22)[2025-05-18]. https://arxiv.org/abs/1807.03748.

[17] PARES F,ARIAS-DUART A,GARCIA-GASULLA D,et al. The MAMe Dataset:On the Relevance of High Resolution and Variable Shape Image Properties[J]. Applied Intelligence,2022,52(10):11703-11724.

[18] TIAN Y T,SUZUKI C,CLANUWAT T,et al. KaoKore:A Pre-modern Japanese Art Facial Expression Dataset[C]//International Conference on Computational Creativity. Online:Springer,2020:415-422.

[19] LIAO P Y,LI X Y,LIU X H,et al. The ArtBench Dataset:Benchmarking Generative Models with Artworks[EB/OL]. (2022-06-22)[2025-05-18]. https://arxiv.org/abs/2206.11404.

[20] MILANI F,FRATERNALI P. A Dataset and A Convolutional Model for Iconography Classification in Paintings[J]. Journal on Computing and Cultural Heritage,2021,14(4):1-18.

[21] LOSHCHILOV I,HUTTER F. Decoupled Weight Decay Regularization[C]//International Conference on Learning Representations. New Orleans:ICLR,2019:1-18.

[22] BAO H B,DONG L,PIAO S H,et al. BEiT:BERT Pre-training of Image Transformers[C]//International Conference on Learning Representations. Online:ICLR,2022:1-19.

[23] CHEN X K,DING M Y,WANG X D,et al. Context Autoencoder for Self-supervised Representation Learning[J]. International Journal of Computer Vision,2024,132(1):208-223.

[24] CHEN X L,HE K M. Exploring Simple Siamese Representation Learning[C]//IEEE Conference on Computer Vision and Pattern Recognition. Nashville:IEEE,2021:15750-15758.

[25] ZHANG H,LI F,LIU S L,et al. DINO:DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection[C]//International Conference on Learning Representations. Kigali:ICLR,2023:1-19.

作者简介:

张　嫱　女,(1988—),硕士,讲师。主要研究方向:人工智能技术在艺术图像生成、艺术图像分类中的应用。

何家成　男,(1999—),硕士研究生。主要研究方向:自监督对比学习、掩码学习。

(*通信作者)**孔令军**　男,(1982—),博士,教授。主要研究方向:人工智能与信息科学交叉研究、通信与存储系统中新型信号处理与纠错码技术。