PixCLIP:多层次特征融合的手写汉字骨架提取

梁晓中,王　涛
(华南师范大学 计算机学院,广东 广州 510631)

摘　要:针对现有的手写汉字骨架提取算法存在的汉字骨架提取不完全、笔画交叉处畸变、笔画断裂等问题,提出一种多层次特征融合与多维度上下文信息增强的手写汉字骨架提取算法,记为 PixCLIP。该算法将多模态预训练模型 CLIP(Contrastive Language-Image Pre-training)的视觉 Transformer 分支与 Pix2Pix 条件生成对抗网络进行多层次特征融合,增强模型整体的全局上下文信息提取能力。对 CLIP 使用视觉提示微调技术(VPT),仅需微调极少数额外参数即可增强其在骨架提取任务的表现。提出多维度特征聚合(MDFA)模块,充分促进 CLIP 的全局特征与 Pix2Pix 局部特征之间多维度特征的自适应融合。引入多头注意力机制与卷积块注意力模块(CBAM),在通道和空间维度上抑制冗余信息。基于在线手写汉字数据集,构建手写汉字图像数据集用于实验。实验表明,与现有最优的骨架提取算法相比,该算法在测试数据集与真实手写汉字图像中均能更好地提取出流畅、完整的汉字骨架;在测试数据集上,此模型 F1 值与联合交并比(IoU)分别达到了 85.62% 和 75.45% 。

关键词:骨架提取;条件生成对抗网络;多模态;CLIP 模型;视觉提示微调
中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1673-629X(2025)01-0021-09
doi:10.20165/j.cnki.ISSN1673-629X.2024.0290

PixCLIP:Multi-level Feature Fusion for Handwritten Chinese Character Skeleton Extraction

LIANG Xiao-zhong,WANG Tao
(School of Computer,South China Normal University,Guangzhou 510631,China)

Abstract:To address the issues of incomplete extraction,distortion at stroke intersections,and stroke breakage in existing handwritten Chinese character skeleton extraction algorithms,we propose a novel algorithm named PixCLIP,which leverages multi-level feature fusion and multi-dimensional contextual information enhancement. It integrates the visual Transformer branch of the multi-modal pre-trained model CLIP (Contrastive Language-Image Pre-training) with the Pix2Pix conditional generative adversarial network to perform multi-level feature fusion,thereby enhancing the overall global contextual information extraction capability of the model. By applying Visual Prompt Tuning (VPT) to CLIP,the performance of CLIP on the skeleton extraction task is improved with only minimal adjustment of additional parameters. Additionally,a Multi-Dimensional Feature Aggregation (MDFA) module is introduced to facilitate the adaptive fusion of global features from CLIP with local features from Pix2Pix. The introduction of a multi-head attention mechanism and Convolutional Block Attention Module (CBAM) further suppresses redundant information in the channel and spatial dimensions. A dataset of handwritten Chinese character images was constructed based on an online handwritten Chinese character dataset for experimental purposes. Experimental results demonstrate that compared with existing state-of-the-art skeleton extraction algorithms,the proposed algorithm achieves smoother and more complete skeleton extraction in both test datasets and real handwritten Chinese character images. On the test dataset,this model achieved F1 scores and Intersection over Union (IoU) values of 85.62% and 75.45%,respectively.

Key words:skeleton extraction;conditional generation adversarial network;multimodal;CLIP model;visual prompt tuning

0　引　言

手写汉字骨架的提取,是从手写汉字图像中提取出代表字形结构的单像素宽度细线图,即“骨架”,并保持原图的拓扑结构。其核心目的是将复杂的手写汉字简化为基本的线条结构,去除书写风格和笔画粗细等次要信息,从而为手写汉字的识别、分析和进一步处

收稿日期:2024-05-27　　**修回日期**:2024-09-27
基金项目:国家自然科学基金项目(61772140)
作者简介:梁晓中(2001-),男,硕士研究生,CCF 会员(R5478G),研究方向为图像处理与计算机视觉;通信作者:王　涛(1975-),男,博士,副教授,研究方向为图像处理与计算机视觉。

理提供基础。

现有的骨架提取技术主要分为两大类:基于像素点间关系的算法和基于深度学习的算法。基于像素点间关系的算法进一步细分为邻域像素点法和像素点间距离法。邻域像素点法通过持续删除前景边缘像素点,直至仅剩下中心线稳定不变,以此确定骨架。每个像素点是否删除取决于该像素点的邻域连通性。而像素点间距离法通过计算前景像素点到最近背景的距离,随后采用特定算法筛选出距离较大的像素点,形成最终的骨架。在基于邻域像素点的骨架提取算法中,Zhang-Suen(ZS)细化算法[1]是一种经典的快速并行处理算法,它主要依据每个像素点周围八邻域像素的状态,来决定该像素点是应该被保留还是删除。通过反复执行删除边界像素的操作,图像的前景边缘逐渐被削减,直至形成细化后的骨架。在算法的每一轮操作中,每个像素点的评估都是独立的,可以并行化处理。ZS 算法能够较快速地得到交叉点、拐角完整的骨架,但存在产生冗余像素、笔画分叉等问题。Ma 等[2]提出了 NFPT 算法,该算法结合了 ZS 算法与单次细化算法[3](One-pass Thinning Algorithm,OPTA),并引入了保留和删除模板来评估每个像素点,可以有效缓解骨架点误删除或非骨架点误保留的情况,最终得到的骨架毛刺更少并且为单像素宽度。基于像素点间距离的骨架提取算法也有很多研究,陈泓汉等[4]先使用 ZS 算法得到待修复的汉字骨架,然后通过两阶段检测法进行修复。第一阶段通过像素点间的距离关系筛选出畸变的分叉点,第二阶段则对这些分叉点进行合并。蔡兴泉等[5]通过对汉字图像使用距离变换创建欧氏距离图,将非骨架点视为山地,模拟漫水淹没山地的过程。在这个过程中,前景轮廓逐步收缩,水面交汇点形成骨架特征点集,最后根据特定条件筛选出骨架点,得到毛刺与冗余较少的汉字骨架图像。这些骨架提取算法处理速度快,但应用到手写汉字骨架提取任务时,往往会因为复杂的汉字形状,可变的笔画宽度,大量的笔画交叉点而出现骨架断裂,交叉点畸变,细化不完全等问题。

随着深度学习技术的发展,深度学习算法的准确率与效率有了大幅度的提升,在骨架提取领域的应用也逐渐变得日益广泛。Nguyen[6]使用 U-Net[7]网络结构,并搭载 CBAM[8]模块。使模型关注重要特征,忽略冗余特征,提升了模型特征提取能力。引入多尺度损失函数,模型能够综合评估不同尺度下的骨架提取能力。Wang 等[9]提出一种用于手写汉字骨架提取的全卷积神经网络,引入回归密集的上采样卷积模块来解决骨架断裂问题,但需要一定的后处理才能得到效果较好的骨架。张子珺等[10]提出一种改进的 Pix2Pix[11]条件生成对抗网络提取书法字骨架,使用分层空洞卷积(Hierarchical Atrous Convolutions Merging,HACM)模块使生成器捕捉更完整长距离上下文信息。使用谱归一化[12]稳定鉴别器训练过程,该网络可直接得到效果较好的书法字骨架。Bi 等[13]提出一种应用于汉字笔画分割的 SSGAN 算法,并嵌入注意力机制,可以快速准确地分割出汉字的笔画。

在目前基于深度学习的骨架提取算法中,卷积神经网络(CNN)因其强大的局部特征提取能力而被广泛使用,但它在全局特征提取方面仍存在不足。而手写汉字存在笔画结构复杂,交叉点多,形态多变等特点,这导致了 Pix2Pix 这类 CNN 模型在提取手写汉字的骨架时,容易过度侧重局部细节而忽略整体结构,导致生成的骨架图像出现交叉点处畸变和笔画断裂等问题。

针对这个问题,该文以 Pix2Pix 模型为主体框架,融合多模态预训练模型(Contrastive Language-Image Pre-training,CLIP)[14],构建一种基于多层次特征融合与多维度上下文信息增强的条件生成对抗网络(PixCLIP)。主要贡献有以下四点:

(1)提出将 CLIP 视觉 Transformer(ViT)[15]分支与 Pix2Pix 条件生成对抗网络进行多层次特征融合。聚合各层次的特征,加强模型对全局上下文信息的提取能力。

(2)提出多维度特征聚合(MDFA)模块,充分促进 CLIP 的全局特征与 Pix2Pix 局部特征之间的多维度特征的自适应融合。提高了模型对手写汉字图像细节特征的处理精度。

(3)引入多头注意力机制(MHA)与卷积块注意力模块(CBAM),在通道和空间维度上抑制冗余信息。并对 CLIP 使用视觉提示微调(Visual Prompt Tuning,VPT)[16]技术,仅需微调极少数的额外参数,即可增强 CLIP 在骨架提取任务中的性能。

(4)实验结果表明,该算法在手写汉字骨架测试集与真实手写汉字图像中,各项指标与效果均优于现有最优的骨架提取算法。

1 文中方法

在图像生成任务中,编码器-解码器架构的模型被广泛采用。该文选用 Pix2Pix 网络结构,这是一种为图像到图像的翻译任务特别设计的条件生成对抗网络(conditional Generative Adversarial Network,cGAN)[17]。与传统的生成对抗网络不同,它主要依赖于有监督学习,通过在生成器和鉴别器中加入条件信息,引导模型的训练过程。这种方法减少了生成图像的随机性,提高了生成内容的可控性。Pix2Pix 网络为

执行复杂的图像到图像翻译任务提供了一种有效且可靠的方法,已成为该领域的一种通用框架。

Pix2Pix 网络在训练过程中使用配对的图像数据作为输入,生成器能够根据输入图像 x 生成尽可能接近真实图像 y 的目标图像,鉴别器能够鉴别一张输入图像是由生成器生成的,还是来自真实的数据集。即生成器负责创建逼真的假图像,而鉴别器负责辨别图像真伪。这两者之间的持续对抗和博弈推动了生成器在图像生成能力上不断提升,鉴别器在鉴别能力上也逐渐增强。最终生成器能创造出足以迷惑鉴别器的逼真图像。这一动态互动过程可以用如下目标函数表示:

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{cGAN}}(G,D) + \lambda \mathcal{L}_{\text{L1}}(G) \quad (1)$$

$$\mathcal{L}_{\text{cGAN}}(G,D) = \mathbb{E}_{x,y}[\log_a D(x,y)] + \mathbb{E}_x[\log_a(1 - D(x,G(x)))] \quad (2)$$

$$\mathcal{L}_{\text{L1}}(G) = \mathbb{E}_{x,y}[\| y - G(x) \|_1] \quad (3)$$

式中, G 表示生成器, D 表示鉴别器。L1 距离损失用于鼓励生成的图像在像素级别与目标图像更接近,从而生成更清晰的图像。λ 用于平衡距离损失与对抗损失。

模型整体框架如图 1 所示。生成器包含编码器和解码器两部分。编码器由若干卷积层组成,并通过多头注意力机制(MHA)增强编码能力;解码器同样由若干卷积层组成,并通过卷积块注意力模块(CBAM)增强解码能力。编码器通过残差连接与解码器进行特征融合。此外,生成器引入了 CLIP 图像编码器,并采用视觉提示微调技术(VPT)进行微调。其中间层特征通过多维度特征聚合模块(MDFA)进行处理,并与编码器的特征进行多层次特征融合。输入汉字图像 x 经生成器 G 后生成骨架图 $G(x)$, $G(x)$ 与 x 构成一对 Fake Pair,目标骨架 y 与 x 构成一对 Real Pair,输入鉴别器 D 进行训练。

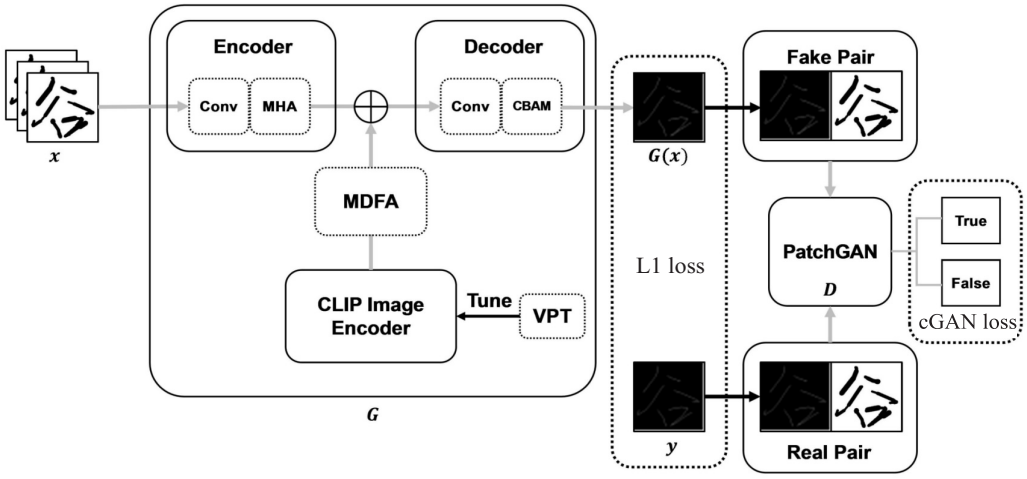

图 1　模型整体框架

1.1　生成器

PixCLIP 的生成器架构如图 2 所示。在编码阶段,生成器进行了 3 次下采样,并引入多头注意力机制使模型能够学习到图像的不同层次特征,加强模型的编码能力。在解码阶段,首先对 CLIP 图像编码器的最后一层输出特征进行解码,然后进行 4 次上采样,通过 MDFA 模块将 CLIP 图像编码器的中间特征与生成器的编码特征进行融合。融合过程的表达式如下:

$$F_{\text{fused}} = \alpha \cdot F_{\text{enc}} + (1 - \alpha) \cdot \text{MDFA}(F_{\text{clip}}) \quad (4)$$

式中, F_{fused} 表示融合后特征, α 是权重系数, F_{enc} 是生成器的编码特征, MDFA 是多维度特征聚合模块, F_{clip} 是 CLIP 图像编码器的中间特征。

随后采用跳跃连接将该融合特征和解码特征进行拼接,恢复图像的细节信息,并搭载 CBAM 卷积块注意力模块增强模型的解码能力。最后通过一个 1 × 1 的卷积层和 Tanh 激活函数得到手写汉字骨架图。

CLIP 模型是 OpenAI 公司提出的多模态视觉-语言模型,用于匹配图像和文本的预训练神经网络,是近年来多模态研究领域的经典之作。CLIP 有两个编码器,分别是图像编码器(ViT 或 ResNet[18])和文本编码器(Transformer)[19]。CLIP 模型通过对比学习[20]方法,在互联网上收集的 4 亿图文对数据集上进行预训练,实现了图像与文本表示空间的有效对齐。得益于在广泛的数据集上进行预训练,CLIP 的图像编码器具备了强大的特征提取能力。因此,该文将 CLIP 的图像编码器融合到生成器中,以增强生成器的特征提取和表示能力。实验证明,将 CLIP 融合到 Pix2Pix 显著提升了模型的整体性能。

该文采用的 CLIP 图像编码器为 ViT-L/14,其架构如图 2 左侧所示。该图像编码器处理分辨率为 224×224 的输入图像,通过 $N = 24$ 个 Transformer 层进行特征提取。ViT-L/14 中的“14”表示图像在输入 ViT 之前,会先被分割成多个分辨率为 14 × 14 图像块,因此

每张图像会被切分为 $m = 256$ 个图像块 I_j。随后每个图像块经过嵌入(Embed)层后转化为一个 $d = 1\ 024$ 维的嵌入表示 e_0^j,并与相应的位置信息结合,该过程可表示如下:

$$e_0^j = \mathrm{Embed}(I_j), e_0^j \in \mathbb{R}^d, j = 1, 2, \cdots, m \tag{5}$$

设 $E_i = \{e_i^j \in \mathbb{R}^d \mid j \in \mathbb{N}, 1 \leqslant j \leqslant m\}$ 为 ViT 第 $i+1$ 层(L_{i+1})的输入。则 ViT 可公式化如下:

$$E_{i+1} = L_{i+1}(E_i), i = 0, 1, \cdots, N-1 \tag{6}$$

其中,对于 ViT 的第 $i+1$ 层(L_{i+1})的处理过程可表示为:

$$Z_i = \mathrm{MultiHead}(\mathrm{LayerNorm}(E_i)) + E_i \tag{7}$$

$$E_{i+1} = \mathrm{MLP}(\mathrm{LayerNorm}(Z_i)) + Z_i \tag{8}$$

其中,Z_i 为 E_i 通过多头自注意力机制后得到的中间表示。

为了图示的简洁性和清晰度,图 2 左侧只展示了 4 个图像块作为例子,并忽略了位置信息的结合。

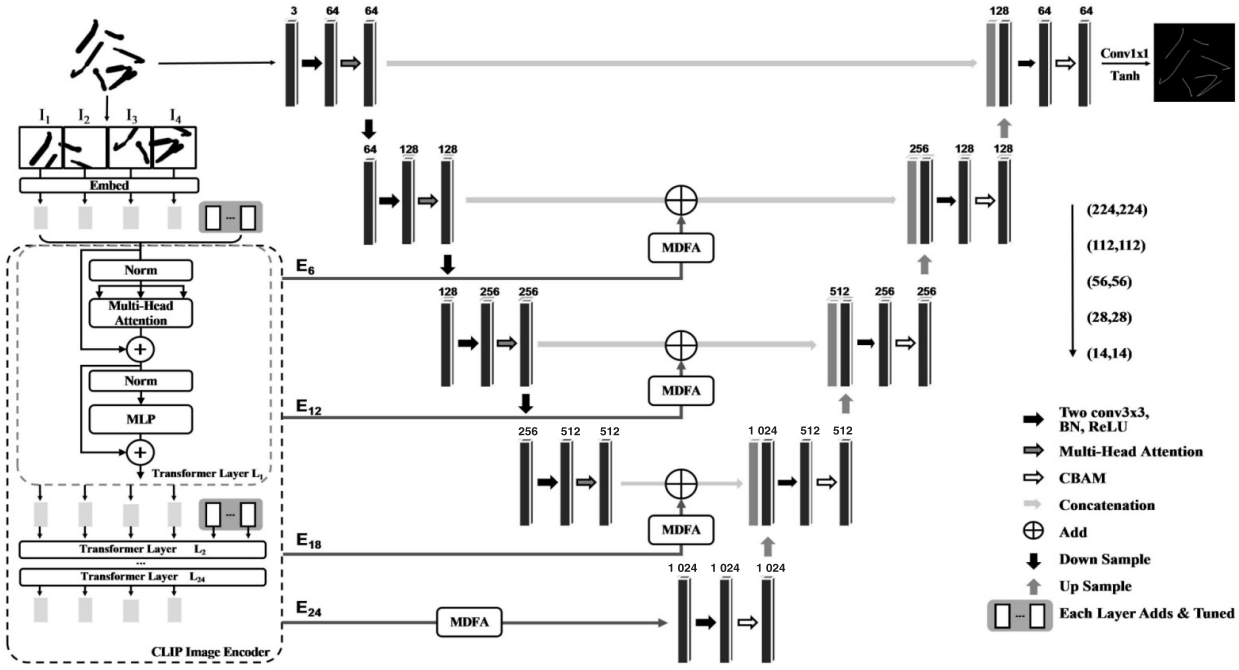

图 2 生成器架构

1.2 视觉提示微调

尽管 CLIP 模型已通过大规模数据集的预训练获得了强大的特征提取能力,但为了进一步增强其在手写汉字骨架提取任务中的表现,该文对 CLIP 的图像编码器进行视觉提示微调。如图 2 左侧所示,对于 CLIP 图像编码器的每个 Transformer 层 L_{i+1} 输入前,与原始图像块嵌入 E_i 并排放置 p 个可学习(Tuned)的向量 p_i^k,均为 $d = 1\ 024$ 维度。则 ViT 可公式化如下:

$$P_i = \{p_i^k \in \mathbb{R}^d \mid k \in \mathbb{N}, 1 \leqslant k \leqslant p\} \tag{9}$$

$$[E_{i+1}, T_{i+1}] = L_{i+1}([E_i, P_i]), i = 0, 1, \cdots, N-1 \tag{10}$$

其中,T_{i+1} 为 P_i 经过 L_{i+1} 后的输出。

训练过程中只对 P_i 的参数进行更新,而冻结 CLIP 图像编码器的其他参数(L_i 中的参数)。通过视觉提示微调,每个层都能够适应性地调整其处理信息的方式,使得 CLIP 能更加精细地针对骨架提取任务调整其特征提取能力。此外,该方法所需要微调的参数量非常少。以 ViT-L/14 为例,其参数量约为 304 M,若每层使用 $p = 20$ 个可学习的提示向量,额外参数量为 $N \times p \times d = 24 \times 20 \times 1\ 024 = 0.49$ M。仅占 ViT-L/14 总参数量的 0.16%。只需要微调极少的参数,就能显著提升模型的性能,大幅节约计算资源,增强了模型对特定任务的适应性。

1.3 多层次特征融合

在 CLIP 图像编码器的 24 层中,将第 6、12、18、24 层输出的嵌入表示(E_6、E_{12}、E_{18}、E_{24})通过多维度特征聚合(MDFA)模块转化成特征图。其中 E_6、E_{12}、E_{18} 与 Pix2Pix 编码阶段的三次下采样特征进行融合,再将融合后的特征通过跳跃连接与解码阶段的特征进行拼接。E_{24} 则直接解码,并进行上采样,与编码阶段特征进行拼接。此过程有效地结合了局部信息和全局信息,提升了模型在手写汉字骨架提取任务上的性能。

在用于特征融合的 E_i 的选择上,需要平衡好 CLIP 与 Pix2Pix 在浅层与深层特征之间的差异,以优化融合效果。若仅使用 CLIP 的深层特征进行融合,虽然能够提供丰富的全局信息和抽象信息,但 Pix2Pix 可能难以理解这些深层特征。反之,如果完全依赖 CLIP 的浅层特征,浅层特征主要包含低级视觉信息,缺乏足够的上下文和全局视角,会限制模型在全局特征提取上的能力。因此,该文选择融合 CLIP 各个层次的特

征,确保模型既能够捕捉到图像的细节信息,也能理解到图像的全局结构,从而提高整体的骨架提取性能。实验证明,综合选择包括浅层、中间层、深层,即多层次的 CLIP 特征,可以最大化模型的性能。

1.4　多维度特征聚合(MDFA)模块

图 3 为 MDFA 的结构。

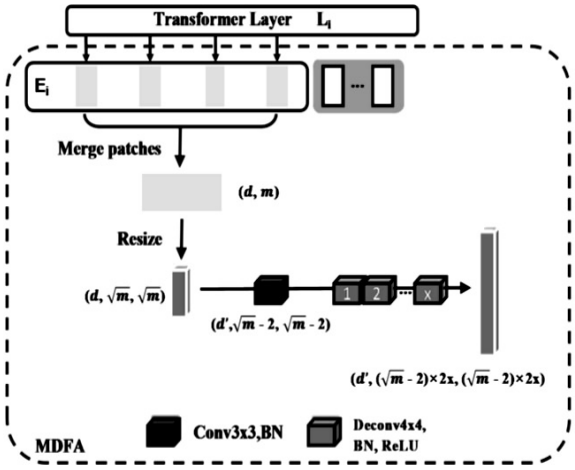

图 3　MDFA 模块

经过 Transformer 第 i 层(L_i)处理后,每个分割的图像块均表示为一个 $d = 1\ 024$ 维的嵌入表示 e_i^j ,共有 $m = 256$ 个嵌入表示。MDFA 首先将这 m 个嵌入表示进行拼接,形成一个尺寸为 $d \times m$ 的特征图,再重塑为 $d \times \sqrt{m} \times \sqrt{m}$ 的特征图 F_r 。该过程表示如下:

$$F_r = R(C(e_i^1, e_i^2, \cdots, e_i^m)), m = 256 \tag{11}$$

其中, C 表示拼接操作, R 表示重塑操作。

为了与生成器中编码层的不同层次特征图进行多维度特征融合,先使用 3×3 的卷积 W_c 将特征图下采样,再用若干个 4×4 的反卷积 W_d 上采样,最终得到的特征图 F 在尺寸和维度上与编码层的特征图一致,并且进一步提取了局部特征,可以进行特征融合操作,该过程表示如下:

$$F = W_d(W_c(F_r)) \tag{12}$$

该模块实现了多维度特征的自适应融合,将 CLIP 模型的全局特征与 Pix2Pix 模型的局部特征进行多维度融合,使模型能够同时捕捉全局上下文信息和局部细节特征,从而提升骨架提取的准确性。通过卷积和反卷积操作,MDFA 将不同来源的特征图尺寸进行匹配,使得特征融合过程更加平滑,并且简单有效。通过合理的下采样和上采样操作,减少了在特征融合过程中不必要的计算开销,提高了模型的运行效率与整体准确率。

1.5　多头注意力机制

多头注意力机制(Multi-Head Attention,MHA)用于融合跨维度信息并实现注意力信息交互,可以显著提升模型的特征提取能力。输入特征图 F 首先经过 3 个并行的 3×3 卷积层 W_1、W_2、W_3,每个卷积层负责捕捉输入特征的不同方面,即不同的“注意力头”。此外,使用一个 1×1 的卷积层 W_s 生成注意力权重 S ,并通过 Softmax 层在通道维度上归一化。最后使用注意力权重 S 对三个卷积层的输出进行加权求和,并与 F 进行残差连接,得到增强特征 F_e 。该过程表示如下:

$$F_1 = W_1(F), F_2 = W_2(F), F_3 = W_3(F) \tag{13}$$

$$S = \mathrm{softmax}(W_s(F)) \tag{14}$$

$$A = \sum_{i=1}^{3} S_i \cdot F_i \tag{15}$$

$$F_e = F + A \tag{16}$$

其中, S_i 表示第 i 个通道的注意力权重, · 表示逐元素相乘操作。

1.6　卷积块注意力模块

CBAM 是一种注意力机制,它可以有效提升模型的整体性能。给定一张特征图 F ,CBAM 模块可以序列化地产生“通道”和“空间”两个维度上的注意力图,提升模型的特征表示能力。CBAM 的结构如图 4 所示,首先将输入特征图通过通道注意力模块进行处理,然后将得到的特征图再通过空间注意力模块进行处理。

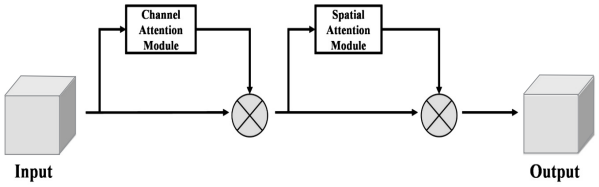

图 4　CBAM

其中,通道注意力模块与空间注意力模块可以分别公式化如下:

$$\mathcal{L}_c(F) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))) \tag{17}$$

$$\mathcal{L}_s(F') = \sigma(f^{7\times7}(\mathrm{concat}[\mathrm{AvgPool}(F'), \mathrm{MaxPool}(F')])) \tag{18}$$

通过加入 CBAM,模型在解码过程中会充分利用通道信息,并在空间通道上抑制冗余信息。

1.7　鉴别器

该文使用的鉴别器为文献[10]中加入了谱归一化的 PatchGAN。以往的鉴别器都是直接将输入图像映射为一个实数,表示其为真样本的概率,关注全局特征和图像的整体布局。而 PatchGAN 将输入图像映射为一个 $N \times N$ 的矩阵,表示每个图像块为真样本的概率,能够集中注意力于图像的局部区域,从而更精细地判断这些局部区域内骨架的真实性。用谱归一化层替代卷积层后的归一化层在一些研究[21]中已经证明可以稳定 GAN 的训练过程,平滑训练过程中的损失。

2 实 验

2.1 数据集

对于手写汉字骨架提取任务,目前尚未有公开的数据集。但在 CASIA 公开的在线手写汉字样本[22]中,记录了多位书写者用笔书写汉字的过程信息,每个手写汉字的笔画轨迹用 (x,y) 坐标序列进行表示。通过将笔画轨迹的坐标进行连接,笔画宽度基本为 1,可以认为是骨架。对骨架图的笔画进行加粗,将笔画控制在一定范围内的随机宽度,可以更好地模拟真实的手写汉字,由此产生的手写汉字与骨架图(目标骨架)构成一对如图 5 所示的训练样本。为了使模型有更强的鲁棒性,该文将多位书写者的手写汉字书写过程用于合成数据集,共 5 000 对训练样本,1 000 对测试样本,每对样本的手写汉字与目标骨架均为 224 × 224 的尺寸。通过实验证明,该合成数据集有效,能很好地模拟真实手写汉字。

图 5 合成数据集

2.2 实验环境

实验环境为 Python3.8、PyTorch1.13.1、CUDA11.6。使用 Adam 优化器,生成器和鉴别器初始学习率均为 0.000 2,指数衰减速率分别为 0.5 和 0.999,训练 50 个 epoch 后学习率均降低为 0.000 1。设置批处理大小为 8,在 NVIDIA RTX4060Ti GPU 上训练 100 个 epoch。

2.3 评价指标

在汉字骨架提取任务中,选用准确率(Accuracy,Acc)、召回率(Recall,Rec)、精确率(Precision,Prec)、F1 值和联合交并比(Intersection over Union,IoU)作为评价指标,可以反映模型在预测精度和鲁棒性方面的性能。

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{19}$$

$$\text{Recall} = \frac{T_P}{T_P + F_N} \tag{20}$$

$$\text{Precision} = \frac{T_P}{T_P + F_P} \tag{21}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{22}$$

$$\text{IoU} = \frac{T_P}{F_P + T_P + F_N} \tag{23}$$

式中,T_P 是实际为正例且预测为正例的数量,F_P 是实际为负例但预测为正例的数量,F_N 是实际为正例但预测为负例的数量,T_N 是实际为负例且预测为负例的数量。

F1 综合考虑了模型不仅要识别出尽可能多的骨架点(高召回率),同时又不能有太多错误的识别(高精确率),其值越接近 1 说明模型效果越好。IoU 衡量了预测骨架与真实骨架的重叠程度,可以用来量化模型预测的骨架与实际手写汉字骨架之间的吻合度。IoU 值越接近 1 说明模型的预测与真实情况越一致。

2.4 对比实验

该文首先对特征融合时的 α 进行对比实验:固定视觉提示向量的数量为 20,调整 α 的值进行定量比较。当 $\alpha = 1$ 时,表示没有进行特征融合。实验结果如表 1 所示。

表 1 不同 α 的对比

α	F1/%	Acc/%	Rec/%	Prec/%	IoU/%
1.00	83.196	99.315	92.485	90.477	71.867
0.90	85.199	99.385	93.227	91.722	74.731
0.85	**85.628**	**99.406**	**93.494**	**91.903**	**75.456**
0.80	84.417	99.357	92.984	91.210	73.594
0.70	84.797	99.375	93.109	91.456	74.227
0.60	83.817	99.338	92.642	90.928	72.782
0.50	83.619	99.340	92.571	90.874	72.747

从表中可以看出,当不采用特征融合时,各项指标均远低于有特征融合,表明融合 CLIP 图像编码器特征的有效性。随着 α 逐渐降低,CLIP 图像编码器的特征权重增加,各项指标逐步增加,当 $\alpha = 0.85$ 时达到最大,随后逐步下降。

该文对 CLIP 图像编码器中的可学习视觉提示向量的数量进行了对比实验:固定 $\alpha = 0.85$ 进行定量比较。

如表 2 所示,随着提示数量的增加,各项指标均逐步上升,当提示数量为 20 时达到峰值,随后呈现下降的趋势。提示数量为 0,意味着不微调 CLIP 图像编码器,只使用原始预训练参数,此时所有性能指标都低于

进行微调后的结果,表明提示微调有助于 CLIP 图像编码器适应当前的下游任务。

表 2　不同提示向量数量的对比

提示数量	F1/%	Acc/%	Rec/%	Prec/%	IoU/%
0	83.325	99.318	92.485	90.619	72.069
10	84.348	99.358	92.920	91.187	73.555
20	**85.628**	**99.406**	**93.494**	**91.903**	**75.456**
30	85.193	99.385	93.187	91.755	74.730
40	85.432	99.398	93.324	91.883	75.153
50	84.741	99.373	93.060	91.429	74.188
100	85.164	99.388	93.328	91.651	74.754

表 3　不同骨架提取算法在合成测试集上的比较

算法	F1/%	Acc/%	Rec/%	Prec/%	IoU/%	训练/ms	测试/ms
改进 ZS	55.784	98.405	77.829	77.205	39.087	–	126
NFPT	60.761	98.636	81.834	78.472	44.184	–	21.3
SSGAN	73.194	99.052	87.935	85.026	58.272	14.5	12.6
改进 Pix2Pix	83.075	99.405	92.367	90.548	72.091	84.2	55.4
文中算法	**85.628**	**99.406**	**93.494**	**91.903**	**75.456**	85.4	59.8

表 3 对比了文献[4]的改进 ZS 算法、文献[2]的 NFPT 算法、文献[13]的 SSGAN 算法、文献[10]的改进 Pix2Pix 算法、文中算法在合成测试集上的性能与时间消耗(每张图像的平均处理时间)。通过与现有的各种骨架提取算法在合成测试集上进行比较,可以看出,虽然在时间消耗上相比现有最优的骨架提取算法略有增加,但文中算法在性能上的各项指标较其他算法均有显著提升。

如图 6 所示,对比了不同算法在合成测试集上的骨架提取效果。从左到右依次为合成手写汉字图像、目标骨架图像、改进 ZS 算法、NFPT 算法、SSGAN 算法、改进 Pix2Pix 算法、文中算法所提取骨架图像。

图 6　不同算法在合成测试集上的骨架提取结果

如图 7 所示,对比了不同算法在真实手写汉字图像上的骨架提取效果。从左到右依次为真实手写汉字图像、改进 ZS 算法、NFPT 算法、SSGAN 算法、改进 Pix2Pix 算法、文中算法所提取骨架图像。

图 7　不同算法在真实手写汉字图像上的骨架提取结果

由图 6 和图 7 可以看出,文中算法泛化能力好,对合成手写汉字图像与真实手写汉字图像均能够提取出流畅、完整的汉字骨架,可以体现出手写汉字的拓扑结构。而其他算法存在较多的骨架区域缺失,交叉点畸变等问题。

2.5　消融实验

表 4 通过比较实验展示了 CLIP 图像编码器中间层特征与生成器编码阶段特征融合的各项指标,旨在验证特征融合策略的有效性。

结果显示,相比单一使用其中一种特征,两者的结合在所有评价指标上均表现更佳。融合特征后,模型的性能显著增强,表明 CLIP 图像编码器特征对于 Pix2Pix 框架中的特征提取有重要作用。

表 5 对比了文中算法去除 CBAM(标为“–CBAM”)和去除 MHA(标为“–MHA”)模块前后的模

型结构在各项性能指标上的差异,证明了这两个模块的有效性。实验结果表明,引入 CBAM 和 MHA 模块能显著提高模型的整体性能,并且当两者同时使用时,性能增益最大。

在 CLIP 图像编码器的 24 层输出特征中,需要选择 4 层用于特征融合。表 6 展示了不同层次选择的性能指标对比。

表 4 单一特征与融合特征的性能比较

F_{enc}	F_{clip}	F1/%	Acc/%	Rec/%	Prec/%	IoU/%
√		83.196	99.315	92.485	90.477	71.867
	√	60.804	98.476	81.495	78.811	44.376
√	√	**85.628**	**99.406**	**93.494**	**91.903**	**75.456**

表 5 CBAM 和 MHA 模块对模型性能影响的比较

网络结构	F1/%	Acc/%	Rec/%	Prec/%	IoU/%
文中算法	**85.628**	**99.406**	**93.494**	**91.903**	**75.456**
-CBAM	84.551	99.363	93.042	91.258	73.742
-MHA	84.629	99.364	92.843	91.526	74.041
-CBAM-MHA	84.079	99.342	92.773	91.064	73.090

表 6 CLIP 图像编码器不同层次选择对比

层次	F1/%	Acc/%	Rec/%	Prec/%	IoU/%
E_1, E_2, E_3, E_4	83.956	99.345	92.648	91.104	73.133
$E_{11}, E_{12}, E_{13}, E_{14}$	84.203	99.348	92.846	91.107	73.254
$E_{21}, E_{22}, E_{23}, E_{24}$	84.708	99.365	92.959	91.498	74.016
E_1, E_2, E_{23}, E_{24}	84.896	99.374	93.120	91.548	74.328
$E_6, E_{12}, E_{18}, E_{24}$	**85.628**	**99.406**	**93.494**	**91.903**	**75.456**

实验数据显示,当单独使用 CLIP 图像编码器的浅层、中间层或深层特征进行特征融合时,模型的性能指标逐步提升,即深层特征的融合对模型性能的增益最为显著。但这些性能指标仍然低于同时融合浅层与深层特征。并且当模型采用多层次特征融合时,即综合使用浅层、中间层和深层特征,其性能达到了最高。这证明了多层次特征融合在增强模型的局部细节捕捉能力和全局上下文理解能力之间提供了有效的协同效应,从而显著提升整体模型性能。

3 结束语

该文首次在汉字骨架提取领域将多模态预训练模型 CLIP 与 Pix2Pix 条件生成对抗网络相结合,提出了一种基于多层次特征融合与多维度上下文信息增强的 PixCLIP 手写汉字骨架提取算法。该算法结合了 Transformer 全局特征提取和 CNN 局部特征提取的优势,将多层次特征进行融合,并引入了注意力机制,增强了模型的特征提取能力。提出了多维度特征聚合模块(MDFA),实现了 CLIP 图像编码器的中间特征与 Pix2Pix 的编码特征之间的多维度特征融合。使用视觉提示微调技术进一步增强 CLIP 模型在骨架提取任务中的性能。通过大量的对比与消融实验,验证了该方法的有效性,能够有效解决现有方法在提取骨架时存在的交叉点畸变、骨架区域缺失等问题。通过在真实手写汉字图像上进行对比实验,该方法展现了良好的泛化能力,可以提取出不同形态手写汉字的骨架。在未来,将进一步扩展文中数据集以获得泛化能力更强的模型。

参考文献:

[1] ZHANG T Y, SUEN C Y. A fast parallel algorithm for thinning digital patterns[J]. Communications of the ACM, 1984,27(3):236-239.

[2] MA J, REN X, YUREVICH T V. A novel fast iterative parallel thinning algorithm[C]//Proceedings of the 2020 4th international conference on vision, image and signal processing. Bangkok: ACM, 2020:1-5.

[3] CHIN R T, WAN H K, STOVER D L, et al. A one-pass thinning algorithm and its parallel implementation[J]. Computer Vision, Graphics, and Image Processing, 1987, 40(1): 30-40.

[4] 陈泓汉,王 涛,熊显航. 二阶段手写汉字骨架提取优化[J]. 计算机技术与发展,2023,33(7):41-46.

[5] 蔡兴泉,杨 哲,蔡润博,等. 基于漫水填充的图像骨架提取方法[J]. 系统仿真学报,2020,32(8):1455-1464.

[6] NGUYEN N H. U-net based skeletonization and bag of

tricks[C]//Proceedings of the IEEE/CVF international conference on computer vision. Montreal: IEEE, 2021: 2105 - 2109.

[7] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]// Medical image computing and computer-assisted intervention - MICCAI 2015: 18th international conference. Munich: Springer, 2015: 234-241.

[8] WOO S, PARK J, LEE J Y, et al. Cbam: convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). Munich: Springer, 2018: 3-19.

[9] WANG T Q, LIU C L. Fully convolutional network based skeletonization for handwritten chinese characters[C]//Proceedings of the AAAI conference on artificial intelligence. [s. l.]: AAAI, 2018: 2540-2547.

[10] 张子珺,陈劲松,钱夕元. 基于改进条件生成对抗网络的书法字骨架提取[J]. 计算机工程,2023,49(10):272-279.

[11] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017: 1125-1134.

[12] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral normalization for generative adversarial networks[J]. arXiv: 1802. 05957, 2018.

[13] BI F, HAN J, TIAN Y, et al. SSGAN: generative adversarial networks for the stroke segmentation of calligraphic characters[J]. The Visual Computer, 2022, 38(7): 2581-2590.

[14] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]//International conference on machine learning. [s. l.]: PMLR, 2021: 8748-8763.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J]. arXiv: 2010. 11929, 2020.

[16] JIA M, TANG L, CHEN B C, et al. Visual prompt tuning [C]//European conference on computer vision. Switzerland: Springer, 2022: 709-727.

[17] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. arXiv: 1411. 1784, 2014.

[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 770-778.

[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st international conference on neural information processing systems. Long Beach: ACM, 2017: 6000-6010.

[20] MO S, SUN Z, LI C. Multi-level contrastive learning for self-supervised vision transformers[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. Waikoloa: IEEE, 2023: 2778-2787.

[21] EGOROV E A, ROGACHEV A I. Adaptive spectral normalization for generative models[C]//Doklady mathematics. Moscow: Pleiades Publishing, 2024: 1-10.

[22] LIU C L, YIN F, WANG D H, et al. CASIA online and offline Chinese handwriting databases[C]//2011 international conference on document analysis and recognition. Beijing: IEEE, 2011: 37-41.