基于草图纹理和形状特征融合的草图识别

张兴园 [1] 黄雅平 [1] 邹 琪 [1] 裴艳婷 [1]

摘 要 人类具有很强的草图识别能力. 然而, 由于草图具有稀疏性和缺少细节的特点, 目前的深度学习模型在草图分类任务上仍然面临挑战. 目前的工作只是将草图看作灰度图像而忽略了不同草图类别间的形状表示差异. 提出一种端到端的手绘草图识别模型, 简称双模型融合网络, 它可以通过相互学习策略获取草图的纹理和形状信息. 具体地, 该模型由 2 个分支组成: 一个分支能够从图像表示 (即原始草图) 中自动提取纹理特征, 另一个分支能够从图形表示 (即基于点的草图) 中自动提取形状特征. 此外, 提出视觉注意一致性损失来度量 2 个分支之间视觉显著图的一致性, 这样可以保证 2 个分支关注相同的判别性区域. 最终将分类损失、类别一致性损失和视觉注意一致性损失结合完成双模型融合网络的优化. 在两个具有挑战性的数据集 TU-Berlin 数据集和 Sketchy 数据集上进行草图分类实验, 评估结果说明了双模型融合网络显著优于基准方法并达到最佳性能.

关键词 草图分类, 注意力机制, 互学习策略, 图像识别

引用格式 张兴园, 黄雅平, 邹琪, 裴艳婷. 基于草图纹理和形状特征融合的草图识别. 自动化学报, 2022, **48**(9): 2223−2232

DOI 10.16383/j.aas.c200070

Texture and Shape Feature Fusion Based Sketch Recognition

ZHANG Xing-Yuan[1] HUANG Ya-Ping[1] ZOU Qi[1] PEI Yan-Ting[1]

Abstract Human has a strong ability to recognize hand-drawn sketches. However, state-of-the-art models on sketch classification tasks remain challenging due to the sparse lines and limited details of sketches. Previous deep neural networks treat sketches as general images and ignore the shape representations for different categories. In this paper, we aim to address the problem by an end-to-end hand-drawn sketch recognition model, named dual-model fusion network, which can capture both texture and shape information of sketches via a mutual learning strategy. Specifically, our model is composed of two branches: one branch can automatically extract texture features from an image-based representation, i.e., the raw sketches, and the other branch can obtain shape information from a graph-based representation, i.e., point-based sketches. Moreover, we propose an attention consistency loss to measure the attention heat-map consistency between the two branches, which can simultaneously enable the same concentration of discriminative regions in the two representations. Finally, the proposed dual-model fusion network is optimized by combining classification loss, category consistency loss and attention consistency loss. We conduct extensive experiments on two challenging data sets, TU-Berlin and Sketchy, for sketch classification tasks. Our dual-model fusion network significantly outperforms baselines, and achieves the new state-of-the-art performance.

Key words Sketch classification, attention mechanism, mutual learning strategy, image recognition

Citation Zhang Xing-Yuan, Huang Ya-Ping, Zou Qi, Pei Yan-Ting. Texture and shape feature fusion based sketch recognition. *Acta Automatica Sinica*, 2022, **48**(9): 2223−2232

随着数字设备手机、平板和绘图板的迅速发展, 手绘草图正在成为直观表达用户想法的主要方式之一[1]. 因此, 草图识别在计算机视觉领域蓬勃发展, 其中包括基于草图的图像检索[2], 草图分析[3], 草图分割[4] 和基于草图的图像合成[5−6] 等.

草图识别的目的是物体类别识别, 该任务相比于图像识别更具有挑战性. 主要原因是图像一般表示为稠密像素[7−8], 但是草图缺乏丰富的颜色细节和视觉线索[9−11], 使得草图特征的表示更加困难. 为此, 近些年来国内外很多研究人员致力于草图识别方面的研究. 早期草图识别主要针对 CAD 图和艺术画[12]. 受到最新提出的大型数据集启发[13], Schneider 等[14] 提出了一系列基于手工特征的草图识别方法. 这些方法将草图看作自然图像并利用方向梯度直方图[15]

收稿日期 2020-02-18 录用日期 2020-05-03
Manuscript received February 18, 2020; accepted May 3, 2020
中央高校基本科研业务费专项资金 (2020YJS046), 北京市自然科学基金 (M22022, L211015) 和中国博士后科学基金 (2021M690339) 资助
Supported by Fundamental Research Funds for the Central Universities (2020JYS046), Natural Science Foundation of Beijing (M22022, L211015) and Postdoctoral Science Foundation of China (2021M690339)
本文责任编委 金连文
Recommended by Associate Editor JIN Lian-Wen
1. 北京交通大学计算机与信息技术学院 北京 100044
1. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044

和尺度不变特征变换[16]提取草图特征, 再利用支持向量机 (Support vector machine, SVM)[17]分类器对草图类别进行预测[18]. 然而, 上述方法普遍存在的问题是浅层特征不能充分表达草图. 因此, Yu 等[18]提出了 Sketch-a-Net 网络, 这是第 1 个使用深度学习的草图识别网络, 通过参数学习方式代替手工设计策略提高分类性能. 受到上述研究工作启发[19−31]设计了新的深度学习框架, 使得草图识别的性能首次超过了人类. 近期基于循环神经网络提取草图特征的工作[23−31], 除了考虑草图内在结构还加入了笔画的时序序列, 即考虑草图的笔画顺序来提取草图特征, 由此进一步提升了草图识别性能.

然而, 大多数深度卷积神经网络模型将草图视为自然图像来获得具有判别力的纹理特征, 而没有考虑草图本身所具有的形状信息. 具体来说, 草图在二维空间中以曲线的形式进行信息传递[24], 因此草图具备很好地描述物体几何形状的特性. 但由于不同人绘画技巧和绘制风格的差异, 对同一物体进行描述的草图形状会千差万别, 而传统的特征描述子并不能很好的描述类间和类内的形状差异性. 本文工作的目标是寻找一种更有效的方法来将草图形状信息融合到端到端的神经网络中, 从而使得深度学习网络具有更好的草图识别效果. 为此, 本文提出了一种新颖的基于双分支互学习的深度学习网络, 即双模型融合网络 (Dual-model fusion network, DMF-Net), 以此实现草图纹理信息和形状信息的结合来进行草图识别. 在训练阶段, 第 1 个网络分支输入原始草图, 并使用传统的卷积神经网络提取纹理信息; 第 2 个分支输入草图的采样点集合, 并使用基于图卷积神经网络提取形状信息; 2 个网络使用互引导机制实现联合训练. 测试阶段将训练好的网络分别提取不同特征并将其融合, 然后输入分类器实现最终草图的类别预测.

在提出的双分支融合网络中, 基于损失函数互引导机制实现的相互学习主要由 2 部分组成: 1) 网络的每个分支使用传统的监督分类损失和基于另一个分支分类概率作为后验的模仿损失. 为此使用 2 个概率分布的 Kullback-Leibler (KL) 距离作为类别一致性损失; 2) 网络基于视觉显著图的一致性计算损失. 显然, 当草图使用两种不同的形式进行表示时, 视觉显著图的区域应该相同或相近, 因此网络将视觉一致性定义为原始草图的显著图和基于点表示草图的显著图之间的欧氏距离. 最后, 将分类损失、类别一致性损失和视觉注意一致性损失结合完成网络参数的训练.

本文主要贡献有: 1) 针对草图识别问题, 首次提出用新的双模型融合网络来提取草图的纹理信息和形状信息; 2) 针对双分支网络的互学习问题, 提出了利用视觉注意一致性损失、分类损失和类别一致性损失联合训练的策略; 3) 在 Sketchy 数据集和 TU-Berlin 数据集上进行了实验验证. 实验结果表明, 本文提出的模型在草图分类任务上取得了最好的效果.

本文结构安排如下: 第 1 节详细阐述了基于双分支网络的草图识别算法; 第 2 节阐述了网络的训练和测试细节; 第 3 节通过与已有算法在公开数据集上进行定性和定量比较, 实现了对本文提出方法有效性的验证; 第 4 节总结本文所研究的工作并提出下一步的研究方向.

1 双分支识别网络

双分支识别网络的框架如图 1 所示, 由 2 部分

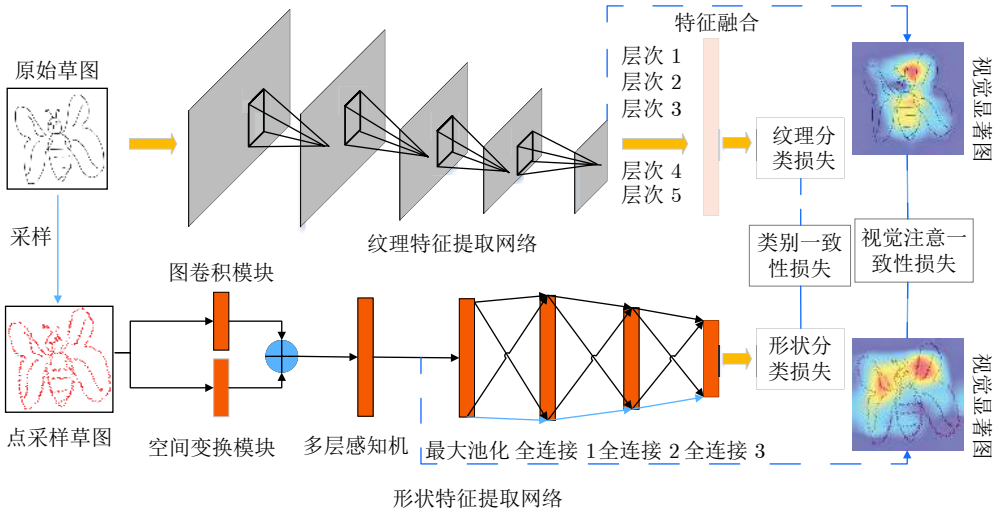


图 1 本文算法总体框架图

Fig. 1 The overall framework of our method

组成: 第 1 部分主要完成基于多层次特征融合网络提取草图的纹理特征, 第 2 部分主要完成基于点采样的图卷积网络提取草图的形状特征, 最后利用分类损失、类别一致性损失和视觉注意一致性损失实现网络参数的优化. 第 1.1 节介绍基于原始草图的神经网络提取纹理特征的过程, 第 1.2 节介绍基于点表示草图的神经网络提取形状特征的过程, 第 1.3 节介绍网络训练使用的损失函数.

1.1　纹理特征提取网络

纹理特征提取网络用于提取草图的纹理信息. 近年来, 基于卷积神经网络提取的信息已被证明能够提供判别性纹理特征以用于草图识别[20−18]. 因此, 本文基于 ImageNet 数据集训练的 ResNet50 模型[25], 提出了纹理特征提取网络 Texture-Net. 为了丰富草图的特征表示, 同时不引入更多的学习参数, 模型利用多层特征融合策略来提取更加丰富的语义信息. 如图 2 所示, ResNet50 网络由 4 个块组成, 每个块的最后 1 个卷积层使用池化操作来保留显著特征并降低特征维度, 然后与网络的第一个全连接层进行特征融合以实现草图纹理特征表示, 即层次 1 到层次 5 的特征融合. 对于第 k 个卷积层所生成的特征图表示为 $F_k \in \mathbf{R}^{H_k \times W_k \times C_k}$, 多层特征图融合用式 (1) 表示为:

$$F_{fusion}(f_i) = [GAP(F_1), GAP(F_2), \cdots, F_{fc}] \quad (1)$$

式中, F_{fc} 表示第 1 个全连接层的特征, $GAP(\cdot)$ 表示在不同特征图上做全局平均池化 (Global average pooling, GAP), 即 $GAP(F_k) \in \mathbf{R}^{C_k}$. 将 GAP 后的卷积层特征与 F_{fc} 特征连接得到最后的草图特征表示. 因此, 最终生成的草图纹理特征包含了从高层语义信息到低层细节信息的多层联合表示. 同时, 网络并没有引入额外的学习参数, 这样保证了融合过程的简单性和高效性.

1.2　形状特征提取网络

与传统神经网络模型对规则网格提取特征的 Texture-Net 不同, 本文提出了基于点集的形状特征提取网络 (Shape-Net) 以对代表性的点集提取结构特征. 两种网络提取特征过程有本质性区别: 对于 Texture-Net, 输入和输出特征图分别表示为 $F_1 \in R_1^{H \times W \times c}$ 和 $F_2 \in R_2^{H' \times W' \times c'}$. 其中 W、H 和 W'、H' 分别表示 F_1 和 F_2 的宽度和高度. c 表示上一层的通道数, c' 表示当前层的通道数. 卷积滤波器表示为一个张量, 即 $F \in R^{s_1 \times s_2 \times c \times c'}$, 其中 s_1 和 s_2 分别表示滤波器的宽度和高度. 与 F_1 相比, F_2 有更低分辨率 ($R_1 < R_2$) 和更多通道 ($c' > c$), 因此更能表示高层语义信息. 上述步骤递归执行最终生成分辨率低但通道数多的特征图. 对于 Shape-Net, 输入是一系列的点并且每个点表示一个特征, 即 $F = \{(p_1, f_1), (p_2, f_2), \cdots, (p_n, f_n)\}$. 网络在输入特征图 F_1 上使用点卷积获得高层语义表示 F_2, 其中 F_2 比 F_1 有更低的空间分辨率 (即更少的点) 和更多的通道数. 此外, F_1 到 F_2 的过程能够将特征聚集成更少点但是每个点包含了更丰富的特征.

以离散点作为网络输入的方法最初用于点云处理[34−27], 主要原因是基于点卷积的神经网络可以大大降低模型的时间复杂度和空间复杂度. Hua 等[26] 构建了针对场景语义分割和目标识别的两个逐点卷积神经网络, 并设计了一种可以输出点云中每一点特征的逐点卷积算子. Wang 等[27] 提出了一个端到端的渐进式学习方法, 即通过训练一个基于图像块的多阶段网络, 在不同的细节级别学习点云信息.

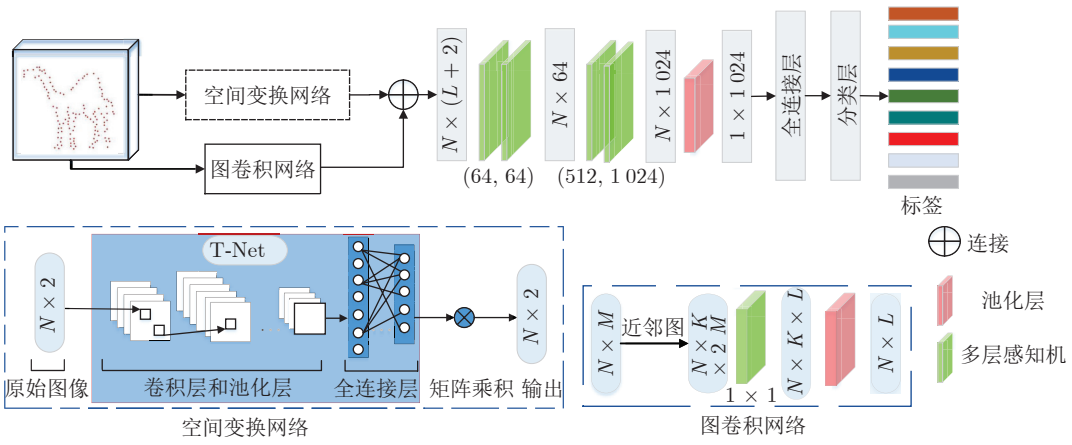


图 2　本文形状特征提取网络的原理框架示意图

Fig. 2　Schematic diagram of Shape-net framework proposed in this paper

基于上述点云中点卷积的思想, 本文通过将草图看成一系列的二维点, 充分考虑草图的空间结构, 最终实现高效、具有鲁棒性的草图特征表示.

通常情况下, 基于点的草图表示具有采样点的起始点位置及草图旋转和平移不变性, 本文采用基于点采样的神经网络作为骨干网络, 该网络模型引入了最大池化和空间变换网络来解决上述两个问题. 但是上述网络设计仍存在不足之处, 即不能获得二维点中局部结构信息, 而邻域信息对特征表达非常重要.

以往的点特征提取方法往往仅利用每个点的特征表示, 而忽略了相邻点的关系, 为了解决上述问题, 人们提出了基于图卷积的神经网络并应用于点云任务中[29−44]. Wang 等[29] 提出使用拉普拉斯算子和池化层次结构动态构件图, 并使用局部谱卷积提取每个点邻域中的结构信息. Simonovsky 等[30] 提出了可以作用在任何图结构上且权重由节点间边权决定的图卷积模型. Kipf 等[44] 提出了一种基于图卷积神经网络的可扩展模型图卷积神经网络 (Graph convolutional network, GCN). 该模型要求网络在训练过程中计算整个图的拉普拉斯矩阵, 然后利用局部谱卷积提取特征, 实验在半监督任务上取得了很好的效果. 基于图卷积思想, 本文提出了基于图卷积的点草图形状特征提取网络. 与 GCN 在谱图上进行卷积操作不同, 此网络在空间域上对图进行卷积操作, 不仅提取当前点的特征, 同时结合了当前点周围的 K 个近邻点特征来表征草图, 即首先以每个点为中心并利用 K 近邻 (K-nearest neighbor, KNN) 算法构建图模型, 然后将中心点特征加入到图中的 K 条边特征中, 再利用卷积操作将 K 个特征整合成一个特征作为中心点最终的特征表示.

下文给出 Shape-Net 的基本结构如下: 第 1.2.1 节通过描述特征提取网络提取草图的形状表示流程, 第 1.2.2 节介绍为了解决草图空间不变性而提出的空间变换网络, 第 1.2.3 节介绍采样点构建邻域图并利用图卷积提取局部结构信息的过程.

1.2.1 特征提取过程

对于 Shape-Net 网络, 首先预处理草图并使用最远采样策略[31] 获取 1024 个采样点. 然后, 利用点卷积学习每个点的特征描述子, 这是提取点特征的关键步骤. 如图 2所示, Shape-Net 实现过程如下: 第 1 步, 将 N 个 M 维采样点输入到二维空间变换网络, 其中 N 表示采样点数, M 表示坐标维度, 由于二维空间坐标为 $(x,\ y)$, 因此 M 取值为 2; 第 2 步, 利用空间变换网络中的 T-Net 模块学习 2×2 仿射矩阵 (第 1.2.2节), 使得网络输入在正则空间中对齐, 同时网络引入图卷积 (第 1.2.3 节) 来从 M 维采样点中学习邻域信息的表示; 最终将图卷积生成的特征图和空间变换生成的特征图拼接生成 $N\times(L+2)$ 维特征图; 第 3 步, 将拼接得到的特征送入多层感知机中, 最终经过多层感知机、最大池化和全连接得到点采样草图的特征表示.

1.2.2 空间变换网络

形状特征提取网络在提取基于点表示草图的信息时, 通过对输入的二维坐标点进行一系列的点坐标变换, 又称为正则变换[36], 来实现形状特征的表达. 而网络对目标的点变换应该具有空间不变性, 即草图旋转和平移不变性. 为此, 网络在进行特征提取之前, 先对点表示的草图进行对齐, 即正则空间对齐, 以保证目标物体对空间变换的不变性. 实际上, 类似于点云模型中使用仿射变换解决几何变形问题[34], 对于基于点坐标的空间变换已经被证明在二维点集匹配[33−35] 和点特征表征[28] 任务中依然重要.

实现对齐操作首先要训练空间变换网络中的 T-Net 模块完成 2×2 仿射矩阵的生成, 然后将训练得到的矩阵与输入的 N 个 2 维采样点进行矩阵相乘来实现最终的对齐目的. 如图 2 所示, T-Net 结构与 Shape-Net 结构类似, 对网络输入的 N 个二维采样点, 分别使用多层感知机、最大池化和全连接层学习特征, 其中全连接层的维度分别设置为 1024、512 和 250, 最后通过矩阵变换得到 2×2 的仿射矩阵.

1.2.3 图卷积网络

邻域点往往具有类似的几何结构[37], 因此通过对每个点构建局部邻域图并对图中的边使用卷积操作, 实现草图局部点集特征的提取. 基本流程如下: 首先, 基于目标点和其 K 个邻近点构建局部图. 然后, 对图中的边利用 1×1 卷积来提取 K 维邻域特征. 将学习到的特征在向前传递过程中, 使用池化层压缩特征以保留重要细节和去掉不相关的信息, 实现局部特征的表达.

具体地, n 个 F 维的点表示为 $\boldsymbol{P}=\{p_1, p_2, \cdots, p_n\}$. 其中, $F=2$, 每个点表示为 2 维坐标. 对于草图中的每个点, 计算无向图 $\xi=(\nu,\varepsilon)$, 其中 $\nu=\{\nu_1, \nu_2, \cdots, \nu_k\}$ 和 $\varepsilon=\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_k\}$ 分别代表顶点和图中的边. 在最简单的情况下, 利用每个点的 K 近邻在二维空间构建无向图 ξ. 提取图中的边特征并表示为 $e_{ij}=h_\theta(x_{mi}, x_j)$, 其中 $h_\theta:\mathbf{R}^F\times\mathbf{R}^F\rightarrow\mathbf{R}^{F'}$ 表示一系列可学习参数 θ 的非线性函数. 最后, 图卷积网络使用池化将学习到的特征进行聚集从而实现每个点局部信息的表达.

1.3 目标函数

网络利用互学习策略进行草图识别模型的训练, 从而进一步提高性能. 本文方法同时训练两个分支, 从而实现网络之间的知识迁移. 网络的目标函数包括分类损失、类别一致性损失和视觉注意一致性损失 3 部分.

1) 分类损失. 网络的两个分支采用 softmax 交叉熵损失函数, 定义如式 (2) 所示:

$$L_c = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{M} 1\left\{y^{(i)}=j\right\}\log_2\frac{\mathrm{e}^{\theta_j^{\mathrm{T}}x^i}}{\sum\limits_{l=1}^{M}\mathrm{e}^{\theta_l^{\mathrm{T}}x^i}}\right] \tag{2}$$

式中, x^i 表示输入的草图, y^i 表示 x^i 对应的标签. N 表示草图的数量. M 表示当前草图类别的数量. θ 表示全连接层的权重, 用来将隐藏层的特征映射为标签.

2) 类别一致性损失. 对于样本 $x_i \in \{x_1, x_2, \cdots, x_N\}$, Texture-Net 利用式 (3) 计算 x_i 属于类 M 的概率:

$$p_1^m(x_i) = \frac{\mathrm{e}^{z_i}}{\sum\limits_{l=1}^{k}\mathrm{e}^{z_i}} \tag{3}$$

式中, z_i 表示隐藏层的输出. Shape-Net 另外一个分支网络使用类似定义并表示为 p_2^m. 为了定量匹配 2 个网络的预测 p_1 和 p_2, 网络使用 KL 散度计算 2 个分布的距离. 从 p_1 到 p_2 的 KL 距离定义如下:

$$\mathrm{D}_{KL}(p_2||p_1) = \sum_{i=1}^{N}\sum_{m=1}^{M} p_2^m(x_i)\log_2\frac{p_2^m(x_i)}{p_1^m(x_i)} \tag{4}$$

3) 视觉注意一致性损失. 通常来讲, 两个分支视觉显著图的一致性能够提高图像分类的性能. 同时, 如果视觉显著图能够反映与类别标签相关的区域, 网络则能够学习到更有判别力的特征. 对于每个原始草图和点采样草图, 网络使用类激活图 (Class activation mapping, CAM)[32] 提取视觉显著图. 具体来讲, 网络在前向激活图上使用加权结合, 然后使用 ReLU 激活函数获得视觉显著图为:

$$H_{ij}^m = ReLU\left(\sum_k\left(\left(\frac{1}{Z}\sum_i\sum_j\frac{\partial y^c}{\partial A_{ij}^k}\right)A^k\right)\right) \tag{5}$$

式中, Z 表示特征图像素数量, A^k 表示卷积层第 k 个特征图.

为了增强显著图的一致性, 网络使用均方误差计算原始草图显著图 H^s 和基于点采样草图视觉图 H^p 之间的损失为:

$$L_a = \frac{1}{NMHW}\sum_{n=1}^{N}\sum_{m=1}^{M}\left[H_{nm}^s - H_{nm}^p\right]_2 \tag{6}$$

式中, H_{nm} 表示第 n 张图像属于第 m 类的视觉显著图.

最终, 将式 (2)、式 (4) 和式 (6) 分别结合实现网络 Texture-Net 和 Shape-Net 的训练, 如下所示:

$$L_{total}^T = L_c^T + D_{KL}(p_2\,\|p_1) + \lambda_1 L_a \tag{7}$$

$$L_{total}^S = L_c^S + D_{KL}(p_1\,\|p_2) + \lambda_2 L_a \tag{8}$$

2 网络的训练和测试

实验使用 Pytorch 深度学习开源工具训练网络模型, 所有实验在两块英伟达 TitanXp 显卡上运行.

2.1 训练阶段

本节基于文献 [38] 的训练策略, 在 ImageNet 数据集上训练 Texture-Net. 同时使用随机梯度下降法进行梯度更新, 每批图像数量为 64. 学习率为 0.1, 错误率稳定时学习率减为原来的 0.1 倍. 衰减率和动量分别为 0.0001 和 0.9. 对于 Shape-Net, 使用 Adam 优化器训练网络, 学习率和每批图像数量分别设置为 0.001 和 64, 每训练 40 轮学习率为原来的 0.5 倍. 衰减率和动量参数设置同 Texture-Net. 式 (7) 和式 (8) 中的超参数设置为 1.

2.2 测试阶段

测试过程中, 首先保持图像纵横比并设置草图的最大尺寸为 256 像素. 然后, 填充边缘像素使得图像大小最终为 256 × 256. 在预测草图类别时, 将两个分支的全连接层特征进行连接, 然后输入到 softmax 分类器中, 得到最终的类别预测.

3 实验结果与分析

为了验证第 3.3 节 DMF-Net 模型在草图分类的有效性, 在 2 个公开数据集 TU-Berlin 数据集和 Sketchy 数据集上进行了实验. 第 3.1 ~ 3.2 节对数据集和评价标准做详细描述, 第 3.4 节通过消融实验分析模型不同模块的性能, 第 3.5 节分析点集采样策略和点数对分类性能的影响.

3.1 数据集介绍

1) TU-Berlin 数据集[13] 是提出的第一个广泛应用于草图识别的手绘草图数据集, 其中包含 250 个草图类别, 例如飞机、日历和长颈鹿. 人类在该数据集所有类别的平均识别率为 73.1%. 由于数据集训练数量的有限性, 采用数据增广方法[18, 32] 进行数据

扩展, 包括水平映射和笔画移除.

2) Sketchy 数据集[39] 是第一个用于细粒度草图检索的大型草图数据集, 包含 75 471 张手绘草图并分成了 125 个类别. 此数据集广泛应用于细粒度草图检索并为图像合成和风格转换提供了可能性.

3.2 评价标准

本文使用和文献 [40−51] 一样的训练策略, 在草图分类任务上定量评估模型性能. 具体来说, 在每个类别上使用平均预测评估分类性能. 然后, 在所有类别上使用平均预测率和其他方法比较分类准确率. 同时引入受试者工作特征曲线 (Receiver operator characteristic curve, ROC)[43] 和曲线下面积 (Area under curve, AUC)[50] 对分类效果进行更好的评价.

3.3 草图分类实验结果

本节展示了 TU-Berlin 数据集在草图分类任务上的性能. 类似于文献 [13], 每个类别在训练集上图像数量的设置分别为 8、40、64 和 72. 数据集中的训练图像是随机选择的, 剩余图像作为测试集. 基准方法分为两类: 传统手工特征和深度特征. 对于手工特征, 选择文献 [14] 中提出的 4 种带有空间金字塔的 FisherVector (FV) 特征表示. 同时采用文献 [13] 提出的映射特征进行草图表示. 对于后者, 对比方法包括 SketchPoint[42]、Alexnet[38]、Network in network (NIN)[45]、视觉几何组网络 (Visual geometry group network, VGGNet)[46]、GoogLeNet[47]、SketchNet[41]、Sketch-a-Net[18]、Cousion Network[49]、Hybrid CNN[48]、Landmark-aware Network (LN)[51] 和 SSDA 方法.

对比结果如表 1 所示. 由表 1 可以看出, DMF-Net 在所有设置都优于其他方法, 并且相对于最优方法有明显提升 (86% 相对 84%、85% 相对 82%、77% 相对 76% 和 60% 相对 59%). 优越性主要得益于纹理特征和形状特征的结合同时引入了 2 个分支的互学习策略进行模型的优化.

图 3 展示了 TU-Berlin 数据集的 6 个类别, 每个类别随机选取 72 张图片进行受试者工作特征曲线的绘制和曲线下面积值的计算. 从图中可以看出, 6 个类别的受试者工作特征曲线均靠近坐标的左上角, 且曲线下面积值均高于随机试验的曲线下面积的值 (0.5), 结果表明了基于互学习的双分支网络在草图分类问题上有着很好的效果.

3.4 消融实验

为了验证本文算法不同模块对分类性能的影响, 模型在 2 个数据集上进行实验测试. 实验包括图卷积、不同的损失函数、2 个分支结构和多层特征融合的影响.

1) 图卷积和不同损失函数的影响. 首先利用实验验证图卷积 (Graph convolution, GC) 对局部几

表 1 不同算法下在 TU-Berlin 数据集上分类准确率的比较 (%)

Table 1 Comparison of sketch classification accuracy with different algorithms on the TU-Berlin dataset (%)

方法	图像数量			
	8	40	64	72
Eitz (KNN hard)	22	33	36	38
Eitz (KNN soft)	26	39	43	44
Eitz (SVM hard)	32	48	53	53
Eitz (SVM soft)	33	50	55	55
FV size 16	39	56	61	62
FV size 16 (SP)	44	60	65	66
FV size 24	41	60	64	65
FV size 24 (SP)	43	62	67	68
SketchPoint	50	68	71	74
AlexNet	55	70	74	75
NIN	51	70	75	75
VGGNet	54	67	75	76
GoogLeNet	52	69	76	77
Sketch-a-Net	58	73	77	78
SketchNet	58	74	77	80
Cousin Network	59	75	78	80
Hybrid CNN	57	75	80	81
LN	58	76	82	82
SSDA	59	76	82	84
DMF-Net	60	77	85	86

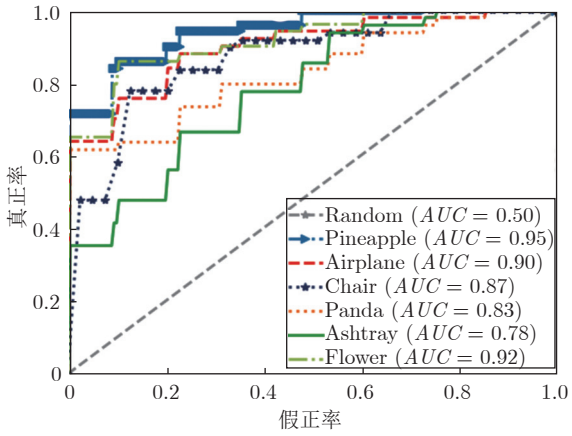


图 3 TU-Berlin 数据集上 6 个类别的受试者工作特征曲线及曲线下面积值

Fig. 3 ROCs and AUC values of 6 classes in the TU-Berlin dataset

何信息具有更强的表示. 然后验证注意力一致性 (Attention consistency, AC) 和类别一致性 (Category consistency, CC) 嵌入模型训练前后对分类准确率的影响. 基础网络 (Base network, BN) 表示在 TU-Berlin 数据集和 Sketchy 数据集上不采用 GC、AC 和 CC 三种策略训练的 Texture-Net 和 Shape-Net.

分类的实验结果如表 2 所示. 通过实验可以得出如下结论: 1) 当结合目标点特征和其临近点特征时, 能够增强局部结构特征的表达并因此解决了不同类别但是结构相似的草图由于局部差异而被错分的问题; 2) 两种草图表示的注意力一致性能够提升性能, 即在 TU-Berlin 数据集上原始模型为 82.71%, 原始模型和注意力一致性结合的准确率为 84.12%. 主要原因是由于视觉注意区域的感知一致性能够帮助两个分支得到更准确的判别区域; 3) 原始模型和类别一致性策略的结合使得模型训练过程中一个分支能够利用其他分支的信息并提升各自的性能; 4) 本文使用原始模型和多个策略的联合, 使得模型能够得到更好的参数优化和特征表达.

表 2　在 TU-Berlin 数据集和 Sketchy 数据集上实现草图分类的网络结构分析 (%)

Table 2　Architecture design analysis for sketch classification on TU-Berlin and Sketchy (%)

方法	TU-Berlin	Sketchy
BN	82.71	85.75
BN + GC	83.93	86.49
BN + AC	84.12	87.07
BN + CC	84.75	87.36
BN + GC + CC	85.47	87.64
BN + AC + CC	85.51	87.71
BN + GC + AC + CC	86.12	88.01

2) 两个分支性能评价. 对于构造复杂和类间相似度高的草图, 传统的神经网络提取的视觉特征往往不能很好的描述草图差异. 因此, 学习基于采样点的结构化特征对草图形状表示具有重要的意义. 为评估两个单独分支和整个网络结构的性能, 在 TU-Berlin 数据集和 Sketchy 数据集上对草图分类准确率进行了实验验证, 所得结果如表 3 所示. 基础网络表示融合 Shape-Net 和 Texture-Net 特征, 而不使用图卷积和互学习损失函数对模型进行优化. 通过对比可以发现, 使用两个分支进行草图的特征表达能够提高分类准确率. 主要原因是纹理特征提取网络对于相似构造的草图差异性具有很差的表达能力. 相反, 形状特征提取网络能获得很好的结构信息, 对草图形状有更好的表达.

表 3　利用双分支神经网络的草图分类准确率 (%)

Table 3　Classification accuracy results using two-branch neural networks (%)

方法	TU-Berlin	Sketchy
纹理网络	81.05	83.18
形状网络	70.87	70.43
基础网络	82.71	85.75

为了进一步验证每个模块对草图分类的重要性, 从 TU-Berlin 数据集中随机挑选了 13 个类别进行定量实验分析. 如图 4 所示, 相比于只使用纹理特征表示草图, 加入形状特征后的草图分类准确率平均提升了 1.5%, 加入形状特征和互学习策略的分类准确率平均提升了 4%. 另外可以发现, 对于一些对于识别困难的样本, 如 Panda、Flying bird 和 Wheel, 无论单独加入形状特征还是融合互学习策略, 分类准确率都有大幅度提升. 因此可以得出以下结论: a) 单纯的纹理表征对难例草图具有很低的判别能力; b) 通过对草图进行点采样, 可以挖掘更多的形状信息以及潜在的草图模式, 从而证明了点表示草图对形状信息具有强表达性; c) 互学习策略可以使得 2 个模型相互学习和共同进步, 从而促进两个网络共同学习到更加具有鲁棒性的特征表示, 使得分类性能得到大幅提升.

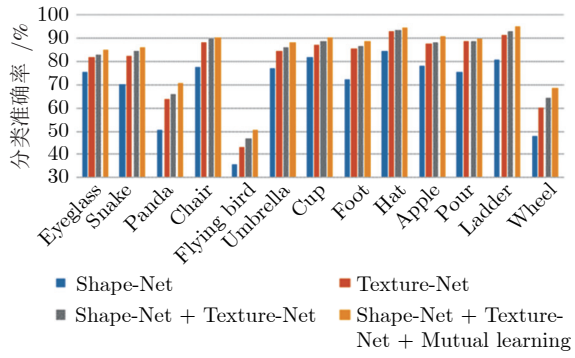


图 4　TU-Berlin 数据集上 13 个类别的分类准确率

Fig. 4　Classification accuracy of 13 classes in the TU-Berlin dataset

3) 多层特征融合评价. 尽管多数神经网络方法只利用全连接层特征表示草图, 但是其他网络层提取的草图特征也能够表达草图并提升分类性能. 因此, 为了融合更多的草图信息, 本文将底层特征和高层特征结合以生成更加丰富的语义信息. 如表 4 所示, {1, 2, 3, 4} 表示图 2 中 ResNet50 第 1 个全连接层融合层次 1、层次 2、层次 3 和层次 4 四个模块中最后一个卷积层经过池化生成的特征表示. 通过对比可以发现, 2 个数据集准确率有相同的趋势:

融合 4 个隐藏层的特征时, 分类准确率最高. 同时, 分类性能逐渐趋于饱和, 并且第 1 层和第 2 层中的卷积层特征对最终草图分类准确率影响很小.

表 4 不同层的分类准确率结果 (%)
Table 4 Classification accuracy results using given feature levels (%)

方法	TU-Berlin	Sketchy
{4}	85.83	87.23
{3, 4}	86.01	87.87
{2, 3, 4}	86.06	87.93
{1, 2, 3, 4}	86.12	88.01

3.5 点采样策略和点数对草图分类的影响

采样常用的策略主要包括均匀采样和随机采样两种. 两种采样结果如图 5 所示, 本文实验采用了均匀采样策略. 通常情况下, 随机采样容易生成抖动噪声, 因此 Shape-Net 模型引入了池化操作和空间变换网络来消除噪声影响. 两种采样对比实验结果如表 5 所示. 由表 5 可以看出, 均匀采样和随机采样在 TU-Berlin 数据集的草图分类准确率分别为 86.12% 和 86.09%. 综上所述, 随机采样性能基本与均匀采样持平, 同时模型对随机采样具有鲁棒性.

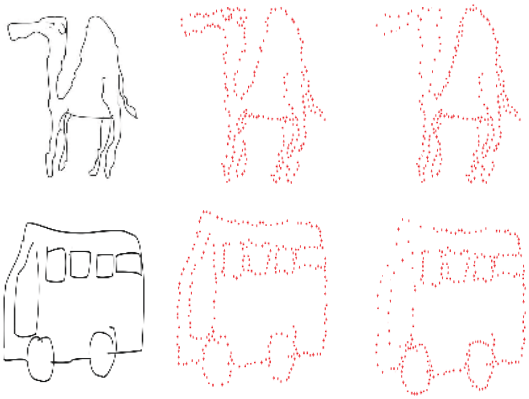


图 5 两种策略在草图点采样的结果示意图
Fig. 5 The point sampling demonstration of two strategies on the sketch

为了验证本文提出的草图识别模型在不同采样点数目下的影响, 在 TU-Berlin 数据集和 Sketchy 数据集上进行了对比实验分析. 从表 6 可以得出以下结论: 1) 随着点数的不断增加, 草图分类准确率在不断的上升. 当采样点数到 1 000 时, 准确率趋于稳定. 同时, 如果采样点过少, 草图分类准确率会大大降低. 因此, 本文设置采样点数为 1 024, 这样在保证计算效率的同时, 准确率也可以达到预期目标;

表 5 2 种采样策略在 TU-Berlin 数据集的分类准确率 (%)
Table 5 Classification accuracy on TU-Berlin dataset using two sampling strategies (%)

方法	分类准确率
均匀采样	86.12
随机采样	86.09

表 6 不同采样点数对分类准确率的影响 (%)
Table 6 Effects of the point number for the classification accuracy (%)

数据集点数	TU-Berlin	Sketchy
32	81.87	83.37
64	82.75	84.35
128	83.23	84.83
256	84.34	85.90
512	85.42	87.36
600	85.75	87.5
750	86.00	88.00
1 024	86.12	88.01
1 200	86.13	88.04
1 300	86.08	88.01

2) 当采样点数继续增加时, 准确率反而会下降. 其主要原因在于, 太多的点数使得点表示的草图集中于表达重复的模式而不是具有判别力的形状, 对于某些难例样本的分类会更加困难.

4 结束语

本文提出了一种基于双分支神经网络结构, 实现手绘草图识别的任务. 网络引入了视觉显著图来学习具有判别力的草图区域, 同时使用注意力一致性和类别一致性的互学习策略实现模型的优化. 在常用两个数据集的实验结果证明了该模型提取的特征优于传统手工特征, 相比于其他几种算法在草图分类任务上拥有更好的表现, 并且对草图的点采样策略具有鲁棒性. 未来将考虑融合笔画顺序到网络中并使用互学习策略, 实现草图性能的进一步优化.

References

1 Huang F, Canny J F, Nichols J. Swire: Sketch-based user interface retrieval. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. Glasgow, UK: 2019. 1–10

2 Dutta A, Akata Z. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 5084–5093

3 Muhammad U R, Yang Y X, Song Y Z, Xiang T, Hospedales T M. Learning deep sketch abstraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA: 2018. 8014–8023

4 Li K, Pang K Y, Song J F, Song Y Z, Xiang T, Hospedales T M, et al. Universal sketch perceptual grouping. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: 2018. 593−609

5 Chen W L, Hays J. SketchyGAN: Towards diverse and realistic sketch to image synthesis. In: Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA: 2018. 9416−9425

6 Zhang M J, Zhang J, Chi Y, Li Y S, Wang N N, Gao X B. Cross-domain face sketch synthesis. *IEEE Access*, 2019, **7**: 98866−98874

7 Pang K Y, Li K, Yang Y X, Zhang H G, Hospedales T M, Xiang T, et al. Generalising fine-grained sketch-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 677−686

8 Fu Xiao, Shen Yuan-Tong, Li Hong-Wei, Cheng Xiao-Mei. A semi-supervised encoder generative adversarial networks model for image classification. *Acta Automatica Sinica*, 2020, **46**(3): 531−539
(付晓, 沈远彤, 李宏伟, 程晓梅. 基于半监督编码生成对抗网络的图像分类模型. 自动化学报, 2020, **46**(3): 531−539)

9 Zheng Y, Yao H X, Sun X S, Zhang S P, Zhao S C, Porikli F. Sketch-specific data augmentation for freehand sketch recognition. *Neurocomputing*, 2021, **456**: 528−539

10 Hayat S, She K, Mateen M, Yu Y. Deep CNN-based features for hand-drawn sketch recognition via transfer learning approach. *International Journal of Advanced Computer Science and Applications*, 2019, **10**(9): 438−448

11 Zhu M, Chen C, Wang N, Tang J, Bao W X. Gradually focused fine-grained sketch-based image retrieval. *PLoS One*, 2019, **14**(5): 217168

12 Jabal M F A, Rahim M S M, Othman N Z S, Jupri Z. A comparative study on extraction and recognition method of CAD data from CAD drawings. In: Proceedings of the International Conference on Information Management and Engineering. Kuala Lumpur, Malaysia: 2009. 709−713

13 Eitz M, Hays J, Alexa M. How do humans sketch objects? *ACM Transactions on Graphics*, 2012, **31**(4): 44

14 Schneider R G, Tuytelaars T. Sketch classification and classification-driven analysis using fisher vectors. *ACM Transactions on Graphics*, 2014, **33**(6): 174

15 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: 2005. 886−893

16 Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, **60**(2): 91−110

17 Chang C C, Lin C J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011, **2**(3): 27

18 Yu Q, Yang Y X, Liu F, Song Y Z, Xiang T, Hospedales T M. Sketch-a-Net: A deep neural network that beats humans. *International Journal of Computer Vision*, 2017, **122**(3): 411−425

19 Bui T, Ribeiro L, Ponti M, Collomosse J. Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. *Computers & Graphics*, 2018, **71**: 77−87

20 Liu Li, Zhao Ling-Jun, Guo Cheng-Yu, Wang Liang, Tang Jun. Texture classification: State-of-the-art methods and prospects. *Acta Automatica Sinica*, 2018, **44**(4): 584−607
(刘丽, 赵凌君, 郭承玉, 王亮, 汤俊. 图像纹理分类方法研究进展和展望. 自动化学报, 2018, **44**(4): 584−607)

21 Xu P, Song Z Y, Yin Q Y, Song Y Z, Wang L. Deep self-supervised representation learning for free-hand sketch. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, **31**(4): 1503−1513

22 Xu P, Huang Y Y, Yuan T T, Pang K Y, Song Y Z, Xiang T, et al. SketchMate: Deep hashing for million-scale human sketch retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA: 2018. 8090−8098

23 Lin Jing-Dong, Wu Xin-Yi, Chai Yi, Yin Hong-Peng. Structure optimization of convolutional neural networks: A survey. *Acta Automatica Sinica*, 2020, **46**(1): 24−37
(林景栋, 吴欣怡, 柴毅, 尹宏鹏. 卷积神经网络结构优化综述. 自动化学报, 2020, **46**(1): 24−37)

24 Liu Y J, Tang K, Joneja A. Sketch-based free-form shape modelling with a fast and stable numerical engine. *Computers & Graphics*, 2005, **29**(5): 771−786

25 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: 2016. 770−778

26 Hua B S, Tran M K, Yeung S K. Pointwise convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA: 2018. 984−993

27 Wang Y F, Wu S H, Huang H, Cohen-Or D, Sorkine-Hornung O. Patch-based progressive 3D point set upsampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 5951−5960

28 Mikolajczyk K, Schmid C. An affine invariant interest point detector. In: Proceedings of the 7th European Conference on Computer Vision. Copenhagen, Denmark: 2002. 128−142

29 Wang C, Samari B, Siddiqi K. Local spectral graph convolution for point set feature learning. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: 2018. 56−71

30 Simonovsky M, Komodakis N. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: 2017. 29−38

31 Eldar Y, Lindenbaum M, Porat M, Zeevi Y Y. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 1997, **6**(9): 1305−1315

32 Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: 2017. 618−626

33 Gold S, Rangarajan A, Lu C P, Pappu S, Mjolsness E. New algorithms for 2D and 3D point matching: Pose estimation and correspondence. *Pattern Recognition*, 1998, **31**(8): 1019−1031

34 Wang Y, Sun Y B, Liu Z W, Sarma S E, Bronstein M M, Solomon J M. Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, 2019, **38**(5): 146

35 Ho J, Yang M H, Rangarajan A, Vemuri B. A new affine registration algorithm for matching 2D point sets. In: Proceedings of the IEEE Workshop on Applications of Computer Vision. Austin, USA: 2007. 25

36 De R, Dutt R, Sukhatme U. Mapping of shape invariant potentials under point canonical transformations. *Journal of Physics A: Mathematical and General*, 1992, **25**(13): 843−850

37 Shen Y R, Feng C, Yang Y Q, Tian D. Mining point cloud local structures by kernel correlation and graph pooling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA: 2018. 4548–4557

38 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: 2012. 1097–1105

39 Sangkloy P, Burnell N, Ham C, James H. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics*, 2016, **35**(4): 119

40 Dey S, Riba P, Dutta A, Lladós J L, Song Y Z. Doodle to search: Practical zero-shot sketch-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 2174–2183

41 Zhang H, Liu S, Zhang C Q, Ren W Q, Wang R, Cao X C. SketchNet: Sketch classification with web images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: 2016. 1105–1113

42 Wang X X, Chen X J, Zha Z J. Sketchpointnet: A compact network for robust sketch recognition. In: Proceedings of the 25th IEEE International Conference on Image Processing. Athens, Greece: 2018. 2994–2998

43 Hanley J A, McNeil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 1982, **143**(1): 29–36

44 Kipf T N, Max W. Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France: 2017.

45 Lin M, Chen Q, Yan S C. Network in network. arXiv preprint, 2013, arXiv: 1312.4400

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

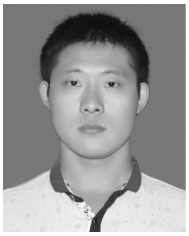
47 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: 2015. 1–9

48 Zhang X Y, Huang Y P, Zou Q, Pei Y T, Zhang R S, Wang S. A hybrid convolutional neural network for sketch recognition. *Pattern Recognition Letters*, 2020, **130**: 73–82

49 Zhang K H, Luo W H, Ma L, Li H D. Cousin network guided sketch recognition via latent attribute warehouse. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: 2019. 9203–9210

50 Lee W H, Gader P D, Wilson J N. Optimizing the area under a receiver operating characteristic curve with application to landmine detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2007, **45**(2): 389–397

51 Zhang H, She P, Liu Y, Gan J H, Cao X C, Foroosh H. Learning structural representations via dynamic object landmarks discovery for sketch recognition and retrieval. *IEEE Transactions on Image Processing*, 2019, **28**(9): 4486–4499

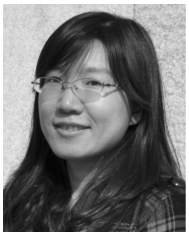


张兴园 北京交通大学计算机与信息技术学院博士研究生. 主要研究方向为深度学习, 数字图像处理和机器学习. E-mail: 15112071@bjtu.edu.cn

(**ZHANG Xing-Yuan** Ph.D. candidate at the School of Computer and Information Technology, Beijing Jiaotong University. His research interest covers deep learning, digital image processing and machine learning.)

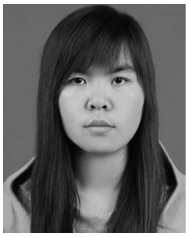
黄雅平 北京交通大学计算机与信息技术学院教授. 主要研究方向为机器学习与认知计算, 人工智能及应用和数字图像处理. 本文通信作者. E-mail: yphuang@bjtu.edu.cn

(**HUANG Ya-Ping** Professor at the School of Computer and Information Technology, Beijing Jiaotong University. Her research interest covers machine learning and cognitive computing, artificial intelligence and application, and digital image processing. Corresponding author of this paper.)



邹 琪 北京交通大学计算机与信息技术学院教授. 主要研究方向为计算机视觉, 人工智能及应用和数字图像处理. E-mail: qzou@bjtu.edu.cn

(**ZOU Qi** Professor at the School of Computer and Information Technology, Beijing Jiaotong University. Her research interest covers computer vision, artificial intelligence and application, and digital image processing.)



裴艳婷 北京交通大学计算机与信息技术学院讲师. 主要研究方向为计算机视觉, 人工智能及应用和数字图像处理. E-mail: ytpei@bjtu.edu.cn

(**PEI Yan-Ting** Lecturer at the School of Computer and Information Technology, Beijing Jiaotong University. Her research interest covers computer vision, artificial intelligence and application, and digital image processing.)