doi: 10.12068/j.issn.1005-3026.2020.07.022

基于生成对抗网络的低分化宫颈癌病理图像分类

李 晨[1], 张家伟[1], 张 昊[1], 汪 茜[2,3]
(1. 东北大学 医学与生物信息工程学院, 辽宁 沈阳 110169; 2. 辽宁省肿瘤医院, 辽宁 沈阳 110042;
3. 中国医科大学附属肿瘤医院, 辽宁 沈阳 110042)

摘 要: 使用生成对抗网络(GAN)扩充宫颈癌病理图像的数据集以提高计算机辅助诊断的准确率. 首先,使用GAN进行细胞质部分图像生成;其次,使用两次 k-means 聚类对生成图像进行筛选;最后,使用Inception-V3模型对数据集进行分类训练. 结果表明,在测试集相同的情况下,该方法可以将总体分类准确率提升约2.5%,尤其对低分化宫颈癌病理图像有显著效果. 通过GAN解决了组织病理学图像无方向性、内容复杂、前景目标规则性差等问题,证明了该方法的有效性及发展潜力.

关 键 词: 宫颈癌辅助诊断;组织病理学图像分类;生成对抗网络;特征提取;k-means 聚类
中图分类号: TP 737.33 **文献标志码**: A **文章编号**: 1005-3026(2020)07-1054-08

Generative Adversarial Networks Based Pathological Images Classification of Poorly Differentiated Cervical Cancer

LI Chen[1], *ZHANG Jia-wei*[1], *ZHANG Hao*[1], *WANG Qian*[2,3]
(1. College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110169, China; 2. Liaoning Cancer Hospital & Institute, Shenyang 110042, China; 3. Cancer Hospital of China Medical University, Shenyang 110042, China. Corresponding author: WANG Qian, E-mail: wangqian_16@163.com)

Abstract: The accuracy of computer-assisted diagnosis can be improved by using generative adversarial networks(GAN) to extend the data set of cervical cancer patholigical images. First, the cytoplasmic part of the histopathological images was generated by GAN; then, k-means clustering was used twice to select images generated by GAN; finally, Inception-V3 model was used to train a classifier. The results showed that the accuracy is improved by an average of 2.5% under the same test data set. Especially, it has significant effect for poorly differentiated cervical cancer pathological images. The non-directionality, complexity of content and poor regularity of foreground target for histopathological images are solved by GAN, which proves the effectiveness and the potential of this method.

Key words: auxiliary diagnosis for cervical cancer; histopathological image classification; generative adversarial networks(GAN); feature extracting; k-means clustering

宫颈癌是最常见的妇科恶性肿瘤之一,其发病率处于全世界女性恶性肿瘤发病率的第三位,因此宫颈癌的预防与诊断工作非常重要[1]. 在宫颈癌的诊断中,组织病理学方法被称为“金标准”,而癌细胞分化阶段的判断是其中的一个关键步骤[2]. 然而,这一步骤需要临床经验丰富的医生对大量组织病理学图像进行分析,其培训成本高、工作强度大,严重阻碍了宫颈癌预防与诊断工作的普及. 因此,使用人工智能方法进行辅助诊断将能够大幅推进以上工作的普及[3-4].

国内在使用计算机对医学图像进行辅助诊断方面的研究起步较晚,最早提到计算机辅助医学图像诊断的是1996年Li等发表的使用计算机对食管癌医学图像的数据分析与处理方案[5]. 在

收稿日期: 2019-07-30
基金项目: 国家自然科学基金资助项目(61806047,81902676); 中央高校基本科研业务费专项资金资助项目(N2019003).
作者简介: 李 晨(1985-),男,辽宁沈阳人,东北大学副教授.

2000 年，Ding 等发表了对宫颈癌细胞涂片计算机自动诊断方案的未来发展意见，说明国内虽然少有系统性的计算机辅助诊断宫颈癌细胞病理图的方案，但逐步意识到癌细胞图像自动诊断的重要意义[6]. 2014 年 Zhao 等把 k – means 颜色聚类用于 CIELab 颜色空间中对宫颈癌细胞进行图像分割[7]. 2016 年 Ning 等使用支持向量机对宫颈组织病理图像分类[8].

国外很早就有将计算机技术应用在细胞病理学显微图像的分类的尝试，Wied 等在 1968 年发表的论文中就提到了他们在所开发的 TICAS 系统中，使用了图像处理的方法对宫颈细胞显微图像中的细胞进行了分类[9]. 近几年利用计算机进行图像分析和处理在国外已有较为广泛的应用[10]. 在卷积神经网络（convolutional neural networks，CNNs）等模型被提出后，陆续发展出很多高效的 CNN 模型，如：VGGNet 与生成对抗网络（generative adversarial networks，GAN）[11－12]. Song 等[13]提出了一种超像素技术和卷积神经网络技术在宫颈癌图像中分割细胞质和细胞核的方法，对细胞核区域检测准确率达到了 94.50%. Purwanti 等[14]提出了一种利用人工神经网络和学习矢量对正常和异常宫颈细胞进行分类的方法，准确率达到了 90%.

目前，对于宫颈癌组织病理学图像计算机辅助诊断系统的研究被专家学者广泛关注，且宫颈癌组织病理学图像较难获取，容易导致训练数据不充分的问题. 本文使用恰当的图像处理与生成方式产生有效的宫颈癌组织病理学图像，扩充训练数据集，提升人工神经网络的学习效率，最终提高人工神经网络对病理图像的识别率，达到了提高辅助诊断效果的目的.

1　宫颈癌组织病理学图像与机器学习

1.1　宫颈癌组织病理学图像

本研究中所使用的宫颈癌组织病理学图像有低分化与高分化两类. 低分化：细胞结构松散，细胞核分布杂乱. 高分化：细胞有一定结构，细胞核分布比较集中. 如图 1 所示，低分化的宫颈癌细胞病变程度严重，从图中已经很难看出细胞原始形态，高分化的宫颈癌细胞病变程度较轻，依然能从图中看出细胞结构[15].

1.2　k – means 聚类算法

k – means 聚类算法具有计算效率高和处理数据量大的优势，其基本思想是：人为给定 k 值，找到即将处理的数据集当中的 k 个聚类点，使所有数据到与它距离最小的聚类点距离的平方总和最小[16].

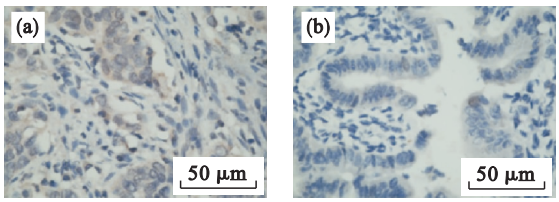

图 1　宫颈癌组织病理学图像实例

Fig. 1　An example of cervical cancer histopathology images

（a）—低分化；（b）—高分化.

1.3　CNN

CNN 是深度学习方法的热点，其基础框架由 3 个部分组成：卷积层，负责对获取的数据进行不同形式的卷积，最终达到提取数据特征并将特征传递给池化层的目的；池化层，接收到卷积层提供的数据特征后进行选择和过滤，精炼数据特征，从而减少神经网络的计算复杂度；全连接层，数据经过卷积层与池化层后会有多个特征传递给全连接层，即全连接层中有多个特征需要分析，最终一般会使用归一化层对所提取的特征进行归一化作为输出[17].

1.4　GAN

GAN 是一种基于博弈论思想开发出来的深度学习模型，主要包含生成模块和判别模块[18]. GAN 的运行原理为生成模块与判别模块不断博弈的过程，生成模块不断生成数据，判别模块学习原始的真实数据，并对生成模块所生成的数据进行判别，而生成模块在生成的过程中依靠判别模块的判别结果来调整自身参数，最终结果为生成模块能够生成判别模块无法判断真伪的数据.

2　使用 GAN 生成宫颈癌组织病理学图像

2.1　研究内容及主要工作

从计算机辅助诊断所使用的图像数据入手，重点研究宫颈癌组织病理学图像的生成方法，主要研究内容如图 2 所示.

步骤 1　因为宫颈癌组织病理学图像较难获取，很容易导致由于数据不足引起的网络训练过程中参数的过拟合现象，所以使用图像切割与仿射变换使数据量扩充为原始数据的 640 倍，对原始的宫颈癌组织病理学图像进行数据扩充.

步骤 2　使用 k – means 进行图像分割，并保

留原始宫颈癌组织病理学图像的主要特征(细胞核部分);使用 GAN 进行训练,并生成次要特征(细胞质与细胞间质部分).

步骤 3　再次使用 k – means 筛选出理想的次要特征生成图,并将其随机拼接为与原始宫颈癌组织病理学图像相同大小的图像,然后与之前保留的细胞核进行融合.

步骤 4　使用现有的 CNN 迁移学习模型对高、低分化宫颈癌图像进行分类训练以及测试,并对网络进行微调使其适用于宫颈癌组织病理学图像的分类任务. 充分利用现有数据和生成数据,合理分配训练数据集与测试数据集,对分类结果进行评价.

步骤 5　在训练与分类过程中,对比使用原始宫颈癌组织病理学图像训练的分类结果和加入生成图像后训练的分类结果并进行分析,以调整生成图像加入的比例,进一步优化分类结果.

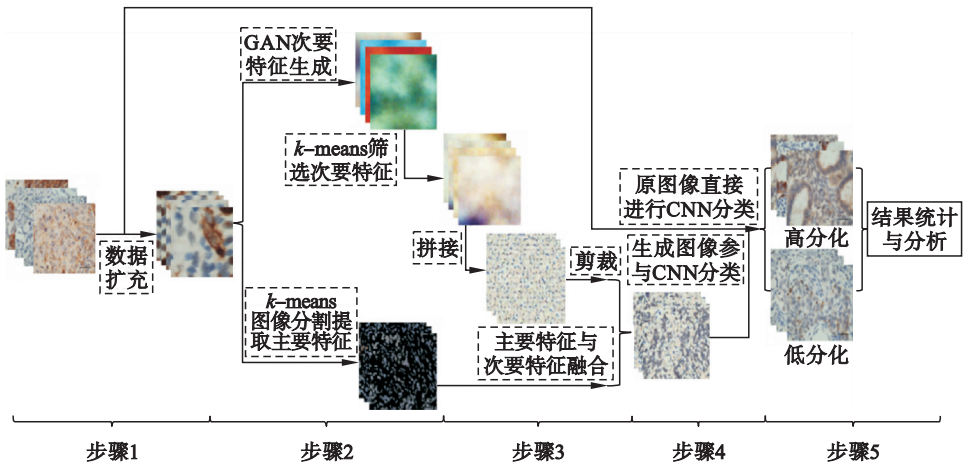

图 2　宫颈癌组织病理学图像生成的工作流程图

Fig. 2　Workflow of cervical cancer histopathology image generation

2.2　数据扩充

在使用计算机进行图像分析或机器学习时,往往需要大量训练数据. 尤其在使用深度学习方法时,数据量不充足很容易导致训练过程中参数的过拟合[19]. 使用合理的方式扩充数据量可以有效地提升实验质量. ①图像剪裁:将原始2 560 × 1 920 像素图像剪裁为 320 × 192 像素,从原本的低分化与高分化各 100 张图像扩充为低分化与高分化各 8 000 张图像. ②仿射变换:因为众多细胞核在显微图像中方向本身是杂乱的,所以旋转变换与翻转变换完全不影响图像的真实性. 因此,本实验中使用旋转变换与翻转变换,将数据量扩充至变换前的 8 倍,达到 64 000 张.

2.3　GAN 的图像生成

由于宫颈癌组织病理学图像在深度学习分类中主要使用细胞核的特征,而细胞核没有固定形状,难以使用 GAN 进行生成,因此本研究使用 GAN 进行图像中次要特征的生成(细胞质与细胞间质部分). 经过合理预处理的宫颈癌组织病理学图像数据有低分化 64 000 张图像与高分化 64 000 张图像. 实验使用 GAN 进行图像生成,训练中将 batch size 设定为 100,一共训练 100 000 迭代次数[19-20]. 实验对生成过程中的结果参数进行分析,迭代次数在 50 000 以上时,判别模块对生成图像的识别正确率为 90% 左右,准确率不再明显上升,生成结果如图 3 所示.

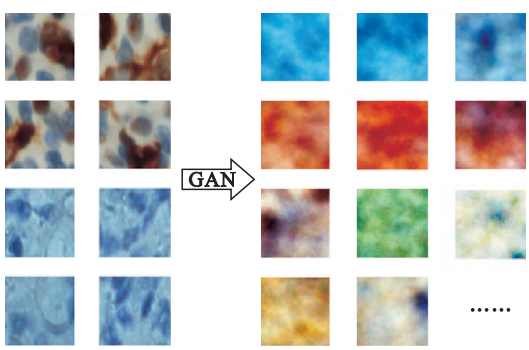

图 3　GAN 生成效果示意图

Fig. 3　Generation effect of GAN

2.4　生成图像的筛选

如图 3 所示,生成图像之间的红绿蓝(red green blue,RGB)值相差很远,很明显有部分图像不适合做数据的次要特征生成. 对此,研究使用 k – means 聚类算法对其进行筛选,其过程如图 4 所示,具体次要特征生成类型见图 5,图 6.

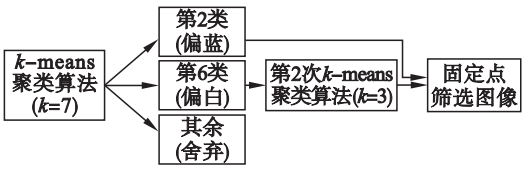

图 4　生成图像筛选过程
Fig. 4　Flowchart of image selection

2.4.1　基于 k – means 聚类算法的图像筛选

生成网络中输出的图像尺寸为 120 × 120 像素，对其图像中全部像素点的 RGB 值求均值，以三维向量形式使用 k – means 算法进行聚类，根据生成图像效果，经过实验，k = 7 时有最好的效果，聚类效果如图 5 所示.

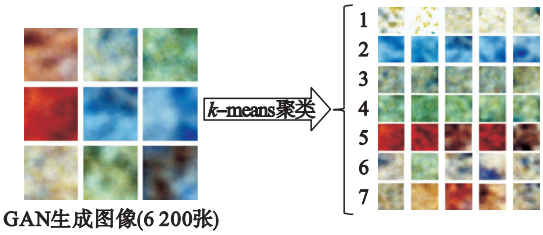

图 5　k – means 算法的图像聚类效果示意图
Fig. 5　Image clustering result by k-means

由图 5 可知，从 7 个聚类结果中筛选出比较适合用于生成组织病理学图像的次要特征的类型：聚类 2 中的结果适合做背景偏蓝的图像的次要特征生成；聚类 6 中的结果适合做背景偏白的图像的次要特征生成；其余的聚类结果与真实图像差距大，故不采用. 其中，聚类 2 中的结果色调偏蓝，而研究中使用的宫颈癌组织病理学图像有相当一部分整体色调偏蓝，因此聚类 2 的整体色调与原图像色调接近，不需要再次对图像的 RGB 值进行 k – means 筛选. 而聚类 6 中结果的 RGB 均值虽然都与背景色调偏白的原宫颈癌组织病理学图像背景接近，但其中还有部分色彩效果稍差的图像，因此需要进行第 2 次 k – means 聚类筛选. 本次聚类设定 k = 3，最终获得正常色调图像的效果如图 6 所示.

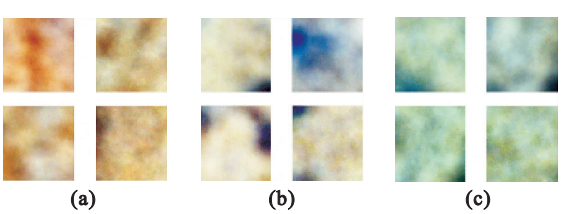

图 6　第 2 次 k – means 聚类效果示意图
Fig. 6　Image clustering result by the second time k-means
(a)—色调偏红；(b)—色调正常；(c)—色调偏绿.

2.4.2　基于 RGB 值的图像筛选

在进行两次 k – means 图像聚类后，生成的图像色彩逐步趋于真实宫颈癌组织病理学图像的色彩，并且有一定量合适的噪声作为生成数据的次要特征. 不过由于 2.4.1 节中的聚类对象是整张图像的 RGB 均值，所以最终聚类结果中难免有一些整体色彩效果不好的图像混入，因此需要对获得的图像进行再次筛选. 观察聚类 2 图像集，可以发现有部分图像中所有像素点的 RGB 值很接近，导致整张图像模糊、纹理辨识度低，不适合做生成图像的背景. 对选取的像素点进行方差计算，将所有像素点的色彩过于接近(方差低)的图像排除，得到纹理特征较强的生成图像，筛选效果如图 7 所示.

图 7　偏蓝图像的 RGB 筛选效果示意图
Fig. 7　RGB selection result of blue images

对聚类 6 中图像的 RGB 值与原图像中背景色偏白的图像细胞质部分的 RGB 值进行对比，将选取的像素点整体色彩表现为偏绿以及轻微偏红的图像都排除，得到最适合做生成数据的次要特征的图像，筛选效果如图 8 所示.

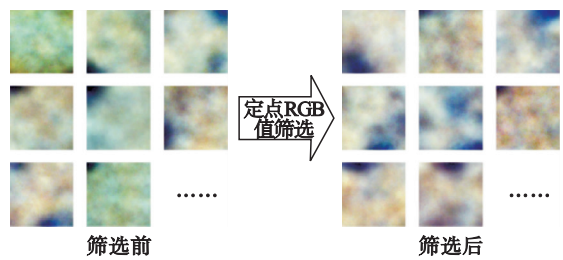

图 8　偏白图像的 RGB 筛选效果示意图
Fig. 8　RGB selection result of white images

2.5　组织病理学图像的生成

2.5.1　组织病理学图像的主要特征(细胞核)的分割

生成的图像数据加入数据集并参与训练，提高准确率的关键在于能否生成或保留图像中的关键特征. 在使用深度学习方法对组织病理学图像的诊断中，细胞核特征是计算机诊断的关键特征，而宫颈癌组织病理学图像中细胞核与细胞质、细胞间质的主要区别在于其颜色特征，因此

本实验中使用 k – means 算法对图像中每个像素的 RGB 值进行聚类,达到细胞核分割的目的,获得并保留宫颈癌组织病理学图像的细胞核部分. 经过测试,本研究设定 $k=3$,其图像分割效果如图 9 所示.

观察图 9a 与图 9b 发现 k – means 算法对像素点 RGB 值的聚类在图像分割中有很好的表现,能将绝大部分细胞核都标记出来,因此使用图 9b 中的标记对图 9a 进行分割,最终得到图 9c 的细胞核分割结果.

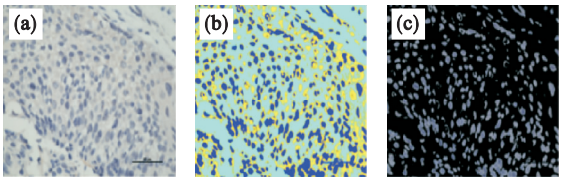

图 9 图像分割结果示意图

Fig. 9 An example of the image segmentation result

(a)—原图像;(b)—k – means 聚类标记;(c)—细胞核分割.

2.5.2 组织病理学图像的次要特征(细胞质)的生成

在通过 GAN 的图像生成及多步筛选之后,得到了一批适合做组织病理学图像次要特征生成的图像. 然后随机选取并将其进行拼接、剪裁,最终获得 2 560 × 1 920 像素的图像,作为所生成的组织病理学图像的次要特征,如图 10 所示.

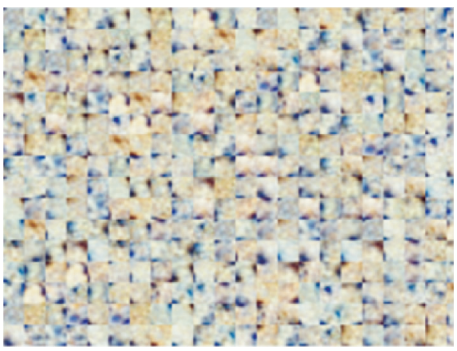

图 10 生成图像次要特征效果示意图

Fig. 10 Generation result of secondary features

2.5.3 主要特征和次要特征的融合

将 2.5.1 节中提取的主要特征与 2.5.2 节中生成的次要特征进行融合. 具体方法为:依据细胞核提取图像的形状将生成的次要特征图像进行分割,分割后与细胞核提取图像叠加生成最终的组织病理学图像,如图 11 所示.

3 实验结果与分析

3.1 组织病理学图像数据的来源

本研究使用低分化与高分化宫颈癌组织病理学图像作为实验数据. 宫颈癌组织切片的制备、显微成像工作以及肿瘤病理类型、分化程度、肿瘤大小的标记工作都由中国医科大学完成.

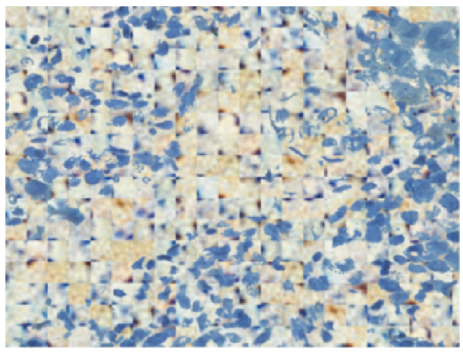

图 11 生成的组织病理学图像

Fig. 11 Generation result for histopathological images

3.2 GAN 的生成实验

使用 320 × 192 像素的原图进行直接生成,将 64 000 张宫颈癌组织病理学图像输入 GAN 进行训练的生成效果如图 12 所示.

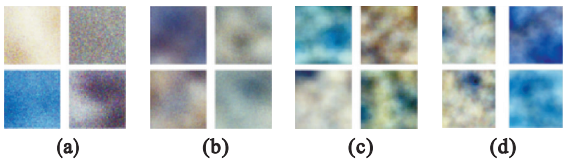

图 12 GAN 生成图像效果示意图

Fig. 12 Generation images by GAN

(a)—迭代次数 = 1 000;(b)—迭代次数 = 10 000;(c)—迭代次数 = 50 000;(d)—迭代次数 = 100 000.

由图 12 可以发现,在 GAN 进行训练的过程中,随着迭代次数的增加,生成的图像纹理越来越清晰,更加接近真实宫颈癌组织病理学图像的纹理效果.

3.3 组织病理学图像的生成

3.3.1 细胞质与细胞间质的生成

在进行筛选工作后,获得与原图十分相近的次要特征(细胞质与细胞间质)图像,如图 13 所示.

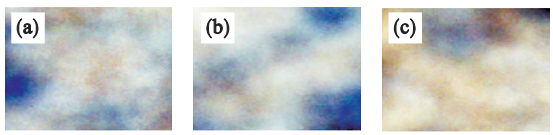

图 13 生成并筛选后的次要特征示意图

Fig. 13 An example of cytoplasm and intercellular after generation and selection

(a)—次要特征实例 1;(b)—次要特征实例 2;(c)—次要特征实例 3.

从图 13 中随机选择 352 张图像进行随机排

序、拼接与剪裁工作，最终生成 2 560 × 1 920 像素的次要特征图像，如图 14 所示.

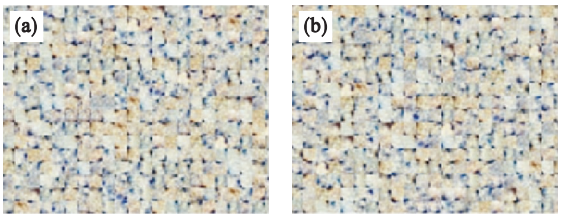

图 14　最终生成的次要特征示例图

Fig. 14　An example of cytoplasm and intercellular after final generation

(a)—次要特征实例 1；(b)—次要特征实例 2.

3.3.2　细胞核的分割

k = 3，4，5 时的细胞核分割效果如图 15 所示.

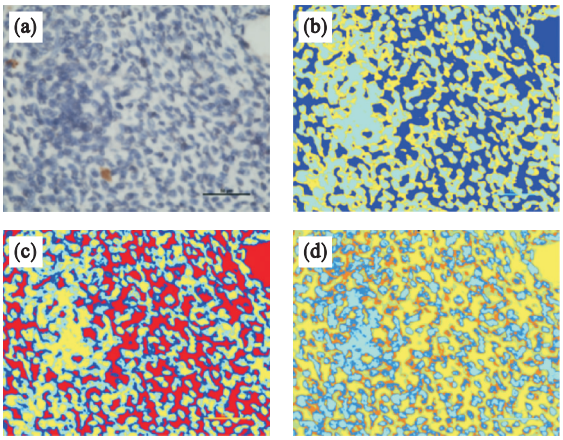

图 15　k – means 算法的聚类标记效果示意图

Fig. 15　Result of cluster marking by k-means

(a)—原图像；(b)—k = 3 聚类效果；(c)—k = 4 聚类效果；(d)—k = 5 聚类效果.

从图中可以看出，在 k = 3 的情况下可以有效地划分出颜色为紫色的区域，细胞核的分割效果较好，如图 16 所示.

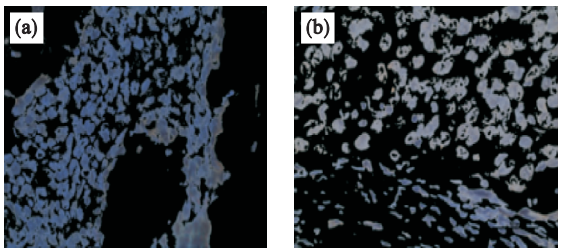

图 16　k – means 算法的图像分割效果示意图

Fig. 16　Result of image segmentation by k-means

(a)—图像分割实例 1；(b)—图像分割实例 2.

3.3.3　细胞核图像与细胞质图像的融合

将 3.3.1 节得到的次要特征的图像与 3.3.2 节得到的主要特征的图像进行融合，生成最终的新图像，如图 17 所示.

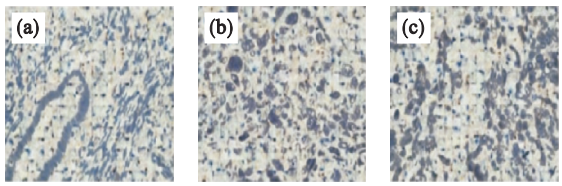

图 17　宫颈癌组织病理学图像生成效果示意图

Fig. 17　An example of the generation result of cervical cancer histopathology images

(a)—生成效果图 1；(b)—生成效果图 2；(c)—生成效果图 3.

3.4　图像分类结果

使用 Inception – V3 进行迁移学习，对高、低分化的宫颈癌组织病理学图像进行 CNN 分类训练，以此来实现辅助诊断的功能. 所生成的 2 560 × 1 920 像素的融合图像，由于分辨率过高，无法直接使用 Inception – V3 网络进行学习，所以先将其统一为 299 × 299 像素的图像，然后再对其进行卷积与池化等操作，最后将图像输入神经网络进行训练. 在训练过程中发现迭代次数在 300 以后，训练曲线波动幅度逐渐变小，图像质量开始趋于稳定，因此在后续实验训练中将迭代次数设置为 300，以保证学习效率. 在此 CNN 参数的设置下，针对每种训练集与测试集的配比进行图像分类测试，并与加入一定比例的生成图像（生成图像占训练集数据的1/3）后的分类结果进行对比. 其中，分类结果为通过分析宫颈癌组织病理学图像判断患者患有高分化宫颈癌还是低分化宫颈癌. 由于本研究是从数据集方面入手尝试提升 CNN 的分类效果，因此实验中合理的数据集的分配比例是必要的. 对训练集、测试集和所加入的生成图像进行统计，最后总结得到实验结果. 实验中训练集与测试集的配比为 5∶5，6∶4，7∶3 和 8∶2，以此测试生成图像数据的可靠性. 结果表明：在 4 种数据集配比情况下，对宫颈癌组织病理学图像进行分类的 4 个指标均有提高：准确率平均上升 1.44%，在8∶2的情况下增幅最大（2.50%）；精确率平均上升 1.82%，在 8∶2 的情况下增幅最大（2.50%）；召回率平均上升 3.18%，在 6∶4 的情况下增幅最大（2.50%）；$F1$ 值平均上升 1.55%，在 6∶4 的情况下增幅最大（2.38%）. 具体结果如图 18 所示.

由图 18 可以发现，在原始宫颈癌组织病理学图像与生成图像数据配比为 5∶5 时，对于低分化图像识别的准确率、召回率及 $F1$ 值均有明显上升；当数据配比为 6∶4 时，对于低分化图像的召回率和 $F1$ 值以及高分化图像的准确率有明显上

升;当数据配比为 7∶3 时,各项指标均有提升,其中对于低分化图像识别准确率的提高较为明显;当数据配比为 8∶2 时,各项指标均有明显提升,增幅较大. 在分类前对图像进行压缩处理,虽然会丢失部分信息,但是并不影响 GAN 生成图像的质量,压缩图像并不会对图像分类结果造成影响. 综上所述,在加入生成图像数据进行训练后,能较稳定地提高其对于低分化宫颈癌组织病理学图像的分类准确率.

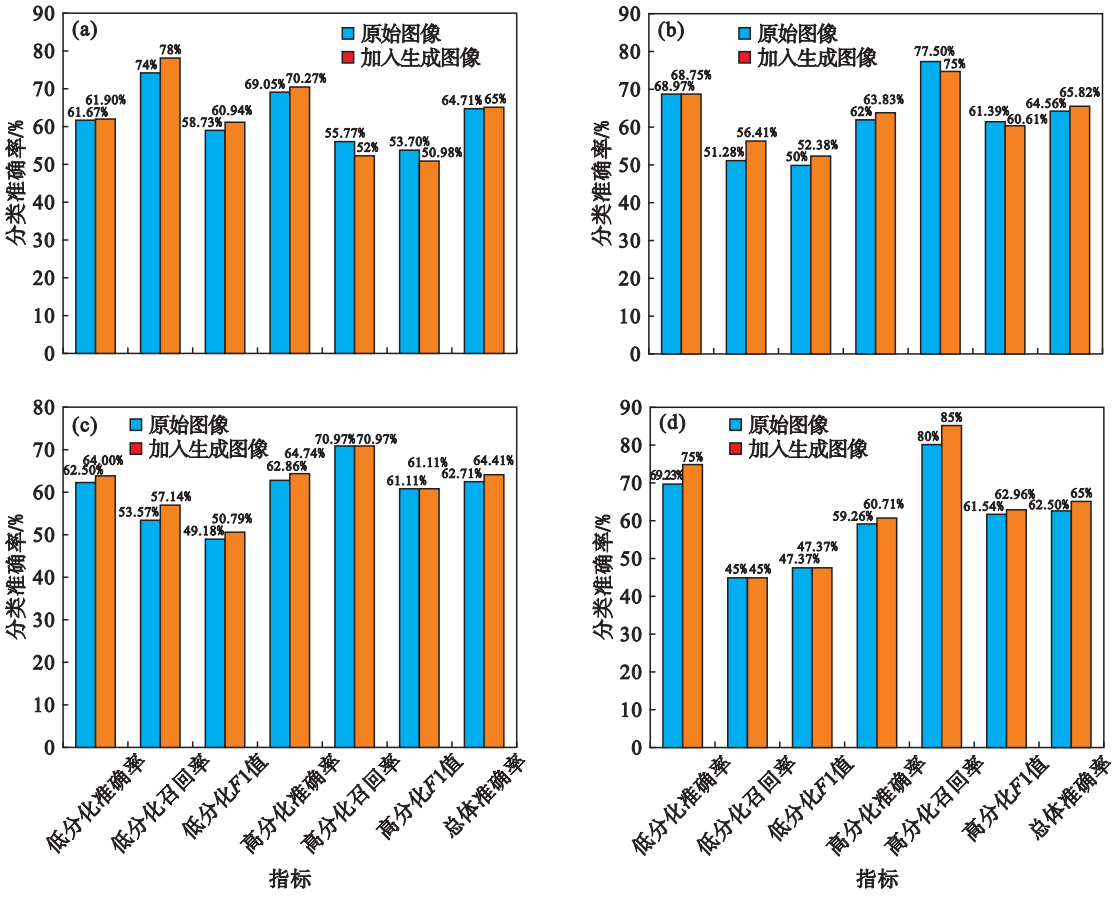

图 18 原始图像与加入生成图像后的训练与分类结果

Fig. 18 The training and classification results of original images and generated images

(a)—训练集:测试集 = 5∶5;(b)—训练集:测试集 = 6∶4;(c)—训练集:测试集 = 7∶3;(d)—训练集:测试集 = 8∶2.

4 结 语

本文以宫颈癌组织病理学图像为研究对象,为计算机辅助诊断提供了一种使用计算机扩充数据量的新方法. 在实验中,通过 CNN 迁移学习的图像分类任务对生成图像的性能进行了验证,发现在训练集中加入生成图像能够有效提高模型对图像的分类效果,尤其对低分化宫颈癌组织病理学图像,识别效果有稳定的提升.

参考文献:

[1] Arbyn M, Castellsague X, De Sanjose S, et al. Worldwide burden of cervical cancer in 2008[J]. *Annlas of Oncology*, 2011, 22(12): 2675 – 2686.

[2] 卞美璐,刘树范. 子宫颈疾病的诊治[M]. 北京:科学技术文献出版社,2001.
(Bian Mei-lu, Liu Shu-fan. Diagnosis and treatment of cervical diseases[M]. Beijing: Scientific and Technical Literature Publishing House, 2001.)

[3] Sukumar P, Gnanamurthy R K. Computer aided detection of cervical cancer using pap smear images based on adaptive neuro fuzzy inference system classifier[J]. *Journal of Medical Imaging & Health Informatics*, 2016, 6(2): 312 – 319.

[4] Loh B C S, Then P H H. Deep learning for cardiac computer-aided diagnosis: benefits, issues & solutions[J]. *Mhealth*, 2017, 3(10): 45.

[5] 李连捷,宋金英. 食管癌计算机辅助诊断中医学图像的量化与数据处理[J]. 河北医科大学学报,1996(1): 23 – 25.
(Li Lian-jie, Song Jin-ying. Quantification and data processing of medical images in the process of computer aided diagnosis for esophageal cancer[J]. *Journal of Hebei Medical University*, 1996(1): 23 – 25.)

(下转第 1064 页)