曾炜峰, 程良伦, 黄国恒. 基于扩散生成的两阶段工业异常检测方法[J]. 广东工业大学学报, 2025, 42(2): 11–19. doi: 10.12052/gdutxb.230204.
Zeng Weifeng, Cheng Lianglun, Huang Guoheng. A two-stage industrial anomaly detection method based on diffusion generative model[J]. Journal of Guangdong University of Technology, 2025, 42(2): 11–19. doi: 10.12052/gdutxb.230204.

基于扩散生成的两阶段工业异常检测方法

曾炜峰，程良伦，黄国恒
（广东工业大学　计算机学院, 广东　广州 510006）

摘要: 现有的工业异常检测方法大多是采用图像修复的思路，在训练阶段利用人工合成伪异常样本的方式，将人工缺陷图像和模型修复后的重建图像进行判别，并计算判别的偏差，得到异常区域。然而，人工合成的缺陷图像与实际缺陷语义关联度不高，不能准确覆盖实际缺陷类型，导致模型鲁棒性不足。为了解决这个问题，本文提出了基于扩散生成的两阶段工业异常检测模型——DADNet(Diffusion Anomaly Detection Network)。首先，利用语义引导的异常生成模块合成已知缺陷，并以此作为异常检测的先验信息。第一阶段采用图像重建模型对工件异常区域进行修复。第二阶段则训练基于修复图像与初始图像的判别模型，用来检测异常区域。此外，本文通过联合关注机制聚合特征，进一步强化重建模型的性能。DADNet在公开的不同材质工件数据集上的性能表现都优于现有模型，并在工业异常缺陷检测中更有应用前景。

关键词: 异常检测；工业检测；图像生成；特征聚合
中图分类号: TP391.4　　**文献标志码:** A　　**文章编号:** 1007–7162(2025)02–0011–09

A Two-stage Industrial Anomaly Detection Method Based on Diffusion Generative Model

Zeng Weifeng, Cheng Lianglun, Huang Guoheng
(School of Computer Science of Technology, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Most existing industrial anomaly detection methods adopt the idea of image restoration, which utilizes artificial synthesis of pseudo anomaly samples in the training stage to discriminate the artificial defect image from the reconstructed images after model restoration, and calculates the deviation of the discrimination to obtain the anomalous region. However, the artificial defect images do not correlate well with the actual defect semantics and cannot accurately cover the actual defect types, resulting in less robustness of the model. In order to solve this problem, this paper proposes a two-stage industrial Diffusion Anomaly Detection network (DADNet). Firstly, a semantically-guided anomaly generation module is utilized to synthesize the known defects, and this is used as the a priori information for anomaly detection. In the first stage, an image reconstruction model is used to repair the abnormal region of the workpiece. In the second stage, a discriminative model based on the repaired image and the initial image is trained and used to detect the anomalous regions. In addition, this paper further enhances the performance of the reconstruction model by aggregating features through the joint attention mechanism. DADNet outperforms the existing models on the publicly available datasets of workpieces with different materials, promising prospect for industrial anomaly defect detection.

Key words: anomaly detection; industrial detection; image generation; feature aggregation

异常检测是一项充满挑战性的任务，而工业异常检测则是其中具有重大意义的领域之一。在工业智能制造的场景中，准确地检测工业零部件表面缺陷，对于整个工业管理和质量控制过程至关重要[1]。

收稿日期： 2023-12-18　　**录用日期：** 2024-05-22
基金项目： 广东省重点领域研发计划项目(2018B010109007)；佛山市重点领域科技攻关项目(2020001006832)
作者简介： 曾炜峰(1998–)，男，硕士研究生，主要研究方向为计算机视觉、图像生成，E-mail：961231790@qq.com
通信作者： 黄国恒(1985–)，男，副教授，博士，主要研究方向为计算机视觉、模式识别和人工智能，E-mail：kevinwong@gdut.edu.cn

然而，在没有专家系统或者先验知识的前提下，工件异常类型和异常区域检测往往是十分困难的。异常检测在训练阶段只包含无缺陷的正常样本，这样既可以避免缺陷、异常样本收集的困难，又可以避免传统方法中的高标注成本与人为标注噪声[2]。因此，开发高效、准确的工业异常检测方法已成为当前学术界和工业界研究的热点之一。

伪异常样本是异常检测任务[3]中的重要一环，它作为异常检测模型的监督信息，影响着模型的训练结果。目前许多伪异常样本的合成方法是通过在正常样本的基础上添加噪声、掩膜等干扰信息，达到模拟缺陷的效果，为模型训练提供有效的监督信息，进一步赋予模型判别异常的能力。然而这类缺陷合成方法使用的干扰信息跟工件无语义关联[4]，会给模型带来大量的冗余纹理信息。在这些冗余信息的干扰下，模型并非获得了工件异常的检测能力，只是获得了一种判别异常数据分布的能力。同时，无关的纹理信息与实际上的工业缺陷差距甚远，由此训练出来的模型无法稳定地识别异常缺陷，这也是目前工业异常检测中的难点之一。此外，在异常检测任务中，特征提取的方式[5]也是影响性能的重要因素，常见的问题是在模型瓶颈层的潜在空间中容易产生特征混淆，从而导致出现大范围的误检结果。

为了解决上述提及的问题，本文采用图像修复的思路，提出一个基于扩散生成的两阶段异常检测模型 (Diffusion Anomaly Detection Network, DADNet)。首先，基于扩散生成模型可控编辑的特点，本文设计出具体缺陷语义的伪异常样本生成方法。同时，DADNet模型的构建分为两个关键阶段：第一阶段是重建模型，其主要任务是对输入样本中的潜在异常区域进行精确修复；第二阶段则是分割模型，它通过分析输入样本与经过修复的样本之间的差异，实现对工业场景中异常的准确检测。具体来说，本文的主要贡献包含以下3个方面。

(1) 提出了面向工业场景的无监督鲁棒异常检测网络，通过语义引导的缺陷生成模块(Semantic Defect Generation Module, SDGM)合成伪异常样本，利用图像生成模型的多样性有效应对工业异常数据稀缺且成本高昂的挑战，包含缺陷语义的异常样本可以为后续的模型训练提供更有效的监督信息，有助于优化正常样本的最小化边界问题，进一步提高异常检测模型的鲁棒性。

(2) 针对以往图像修复模型中出现的正常区域和异常区域重建混淆的问题，提出一个双轨联合的重建特征聚合模块(Dual-track Reconstruction Aggregation Module,DRAM)，聚合潜在空间中的正常样本特征用于重建，进一步加强模型的性能。

(3) 在MVTec数据集上挑选了8个工业场景下的典型工件进行了全面的实验，与最先进的无监督异常检测方法对比，DADNet在不同的类别上都取得最高的得分和最稳定的检测性能。

1 相关工作

1.1 缺陷合成方法

为了更好地实现对缺陷区域的重建，许多研究开始审视基于图像重构方法的训练范式[6]，并得出仅依赖正常样本自身不足以约束出其最小分布边界的结论。因此，合成缺陷伪异常样本对于异常检测任务来说有着重大意义。最广泛使用的一类合成缺陷方法是通过添加随机噪声实现的，如Nakazawa等[7]通过在正常的晶圆图像中加入噪声得到伪异常样本，最终实现了对8种常见的晶圆缺陷的异常检测。此外，还有研究尝试利用数据增强[8]的方式获取伪异常样本图像，进一步提高网络的修复能力。Tayeh等[9]在正常样本中随机擦除任意形状的区域，并用特定的颜色对擦除区域进行填充。然而擦除填充的合成方法[10]忽略了结构信息对图像重建过程的作用，因此网络的修复能力也受到一定的局限。Schluter等[11]通过泊松融合的方式将纹理图像添加在正常图像上来创建伪异常样本。然而，上述人工合成的缺陷缺少与工件相关的语义信息，从而导致在实际应用中性能受限。在一些常见的工业场景下，添加噪声模拟缺陷的方法无法为图像修复模型提供更全面的监督信息，如工件的破损缺陷与结构缺失。

1.2 特征提取方式

对抗生成网络是早期异常检测任务的重要研究方法，其关键是显式或隐式地获得无缺陷数据的特征分布。由于生成模型只生成正常样本，生成的样本与输入的差值则为异常区域。Zhao等[12]通过生成器网络修复样本中的缺陷区域，然后将输入样本和恢复样本进行比较，准确地标识出异常区域。Zhang等[13]提出了DefGAN，通过潜在空间点蚀操作和权重共享来设计重建图像的附加分支，形成新的判别损失以及原始输入图像。此外，随着GAN技术的发展，利用多个GAN在不同特征域之间建立映射变得更加容易。Yu等[14]提出了一种对抗性图像–频率变换网络，应用于道路裂缝的异常检测。然而，工业场景下数据

样本的缺乏也对同时训练两个GAN网络造成阻碍。

目前异常检测中的特征提取方式主要分为基于概率分布和基于先验知识的方法。FastFlow[15]和PaDim[16]属于基于概率分布的特征提取方法，FastFlow利用流模型估计正常样本的二维归一化概率分布，PaDim则是计算目标区域与正常样本的多元高斯分布之间的马氏距离矩阵来评估异常。RD模型[17]和STFPM模型[18]受到知识蒸馏的启发，从预训练模型中提取多尺度特征，前者在训练过程中约束编码器和解码器之间的特征余弦距离，后者旨在利用特征匹配使学生网络接收不同尺度知识的混合，从而允许各种尺寸的异常检测。虽然这些方法提取方式不断创新，但是所提取的瓶颈层特征对于正常样本表征能力依旧欠缺，尤其是表现在缺乏全局特征信息与局部特征信息的聚合。

2　DADNet

2.1　语义引导的缺陷生成模块(SDGM)

以往基于重建的异常检测方法大多采用添加随机噪声模拟异常样本，利用随机噪声训练出来的检测模型毫无疑问可以很好地处理高斯噪声和正常样本之间的特征分布边界问题。然而，在实际检测过程中，缺陷和样本正常区域存在语义信息关联，这类缺陷在特征分布上比随机噪声要更接近正常样本，这也是目前求解正常样本的最小化边界问题中的最大瓶颈。因此，本文中提出了一个语义引导的缺陷生成模块SDGM，通过大型语言模型(Contrastive Language-image Pre-training, CLIP)[19]和文本到图像模型Stable-Diffusion[20]的结合，实现融合语义信息的伪异常样本生成。

如图1所示，SDGM的框架主要包含3条过程分支。首先，训练样本经过随机掩膜生成模块会产生一个尺寸大小为$H\times W$的掩膜$\boldsymbol{M}_{\mathrm{g}}$，该掩膜作为缺陷区域的生成标签。其次，在第二条分支中，正常样本经过预训练的扩散编码器降维，得到潜在图像数据$\boldsymbol{z}_t$，该过程可由式(1)表示。

$$\boldsymbol{z}_t=\sqrt{\overline{\alpha}_t}\boldsymbol{z}_0+\varepsilon\sqrt{1-\overline{\alpha}_t} \tag{1}$$

式中：t为当前潜在数据的扩散步数，$\boldsymbol{z}_0$为输入样本的原始潜在图像表示，$\sqrt{\overline{\alpha}_t}$为$t$步扩散过程的平均图片权重，$\sqrt{1-\overline{\alpha}_t}$为对应噪声的平均权重，$\varepsilon$为一个满足正态分布的随机噪声，$\varepsilon_t\in N(0,1)$。

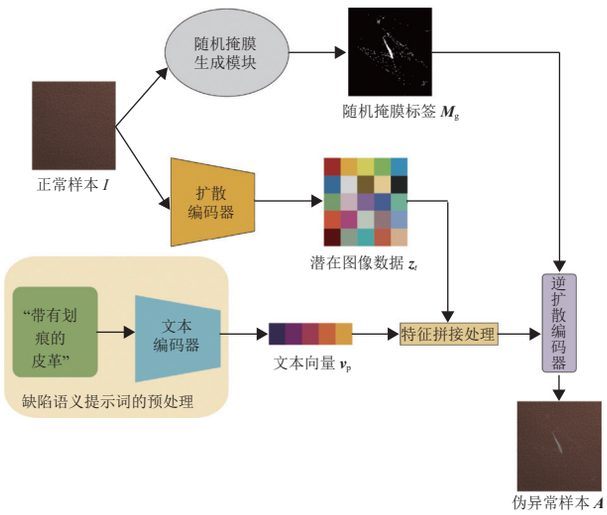


图 1　SDGM的网络结构

Fig.1　The structure of SDGM

经过上述正向扩散过程后，t时刻得到的潜在数据$\boldsymbol{z}_t$会无限逼近高斯噪声，在一些图像生成任务中，往往会直接从噪声图中进行逆扩散操作，从而预测出原图像。然而，为了实现对生成图像的有效语义引导，需要在逆扩散过程中加入条件控制信息。本文采用了CLIP模型对文本信息进行编码，由于CLIP模型在训练过程中会将图像编码和文本编码进行交互训练，因此CLIP模型提取到的文本编码可以直接与图像编码进行相似度计算。在Stable-Diffusion模型的编码器和解码器中存在多个多头注意力，训练过程中会对图像潜在数据和控制图像生成的文本向量$\boldsymbol{v}_{\mathrm{p}}$做相关性计算[21]，计算过程可由式(2)表示。

$$\mathrm{Attentinon}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}}}})\boldsymbol{V} \tag{2}$$

式中：$d_{\boldsymbol{K}}$为矩阵$\boldsymbol{K}$每一列特征的维度，可以通过调节该参数来防止内积计算结果过大。$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$为多头注意力相关性计算中的3个权重矩阵，具体的计算公式为

$$\boldsymbol{Q}=\boldsymbol{W}_{\boldsymbol{Q}}\boldsymbol{z}_t \tag{3}$$

$$\boldsymbol{K}=\boldsymbol{W}_{\boldsymbol{K}}\boldsymbol{v}_p \tag{4}$$

$$\boldsymbol{V}=\boldsymbol{W}_{\boldsymbol{V}}\boldsymbol{v}_p \tag{5}$$

式中：$\boldsymbol{W}_{\boldsymbol{Q}}$、$\boldsymbol{W}_{\boldsymbol{K}}$、$\boldsymbol{W}_{\boldsymbol{V}}$为多头注意力模块的模型参数，它们在训练过程中不断更新，以获取最小化预测误差。在每一轮循环中，Stable-Diffusion都会根据图像潜在数据与文本编码的相关性来更新多头注意力权重矩阵。通过相似度关联的机制，确保Stable-Diffusion模型的逆扩散过程可以实现有效的语义引导生成。

最后一条分支就是对应上述的语义引导的逆扩

散过程，CLIP模型先对输入的文本提示进行编码预处理，将它们转换成文本向量$\boldsymbol{v}_{\mathrm{p}}$。随后把文本向量和正向扩散过程得到的潜在数据拼接在一起，再输入到预训练好的逆扩散编码器中。逆扩散编码器会根据文本向量的引导在原图基础上生成对应的缺陷，最后利用随机掩膜合成出最后的异常样本，过程如式(6)所示。

$$\boldsymbol{A} = I \odot (1 - \boldsymbol{M}_{\mathrm{g}}) + D \odot \boldsymbol{M}_{\mathrm{g}} \tag{6}$$

式中：$\odot$为按像素进行的乘法操作，D为扩散生成的缺陷图像。随机掩膜可以确保语义缺陷在指定的区域生成，并为后续的判别模型提供准确的标签，实现对异常缺陷的精细分割。

2.2 两阶段异常检测网络

基于图像修复的异常检测方法可以分为图像重建和差异检测两个步骤。在图像重建中，样本数据将会被一个可以有效重建正常数据图像表示的自动编码器进行重建。重建后的图像表示会和原始样本经过拼接输入到差异检测的判别分割网络，计算差异并最终检测出异常区域。本文提出的DADNet沿用基于图像修复的异常检测思路，属于两阶段异常检测网络，其中重建网络和判别分割网络均采用U-Net作为编码器解码器的基本结构。

在重建网络中，需要训练模型提取正常样本共有特征并根据所提取特征尽可能重建图像的能力。重建网络的结构如图2所示，正常样本I经过2.1节的SDGM模块得到异常样本$\boldsymbol{A}$，将$\boldsymbol{A}$输入重建网络，经过编码器提取特征和解码器恢复重建，得到重建后图像R。

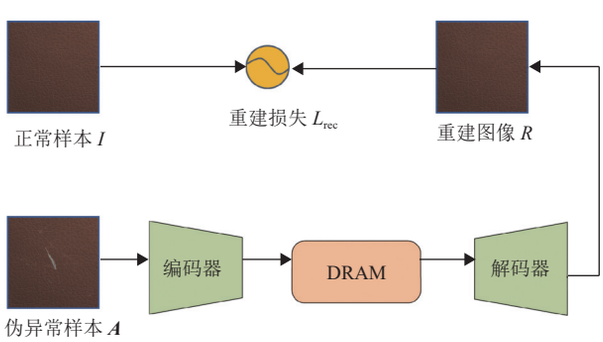


图 2 重建过程的网络结构

Fig.2 The structure of reconstructive network

本文采用了L2损失和结构相似性(Structural Similarity, SSIM)作为损失函数[22]，从而对重建网络进行约束。L2损失基于像素之间相互独立的假设，逐像素进行计算，可以用式(7)来表述。

$$L_{\mathrm{l2}}(I,R) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (I(i,j) - R(i,j))^2 \tag{7}$$

式中：H、W分别为图像的高度和宽度，$I(i,j)$、$R(i,j)$分别为正常样本、重建图像的(i,j)位置对应的像素值。虽然逐像素比较可以考虑到每个像素点间的差异，但是这种约束过强会导致重建图像出现细节的缺失。为了更好地重建细节，加入考虑人类视觉感知的结构相似性作为损失函数的一部分。结构相似性主要从亮度、对比度和结构3个方面评价重建质量，具体的计算可以用式(8)来表述。

$$\mathrm{SSIM}(I,R) = \frac{(2\mu_I\mu_R + C_1)(2\sigma_{IR} + C_2)}{(\mu_I^2 + \mu_R^2 + C_1)(\sigma_I^2 + \sigma_R^2 + C_2)} \tag{8}$$

式中：μ_I、μ_R为样本I、R的平均值，用于作为亮度相似的衡量；σ_I、σ_R为样本I、R的标准差，用于作为对比度相似的衡量；σ_{IR}为样本I、R的协方差，用于作为结构相似程度的衡量。C_1和C_2为正值常数，用于防止公式出现分母为0的异常情况。结构相似性的取值范围是$(0,1)$，当两幅图像相同时取得最大值1。结构相似性的损失函数可由式(9)来表述。

$$L_{\mathrm{SSIM}}(I,R) = 1 - \mathrm{SSIM}(I,R) \tag{9}$$

针对重建网络的损失函数可以用式(10)来表述。

$$L_{\mathrm{rec}} = (1-\lambda)L_{\mathrm{l2}}(I,R) + \lambda L_{\mathrm{SSIM}}(I,R) \tag{10}$$

式中：λ为实验中用于稳定模型训练收敛的超参数，本文中的实验结果是在设置$\lambda = 0.85$下得到。

判别分割网络的结构如图3所示，将异常样本$\boldsymbol{A}$和重建图像R进行通道维度的拼接输入到判别分割网络。它的整体架构和重建网络相似，但是在得到瓶颈层后的解码过程中会加入跳跃连接层，拼接编码过程中相同尺寸的特征。这些浅层特征包含丰富的细节信息，更利于实现精细分割，最终输出得到预

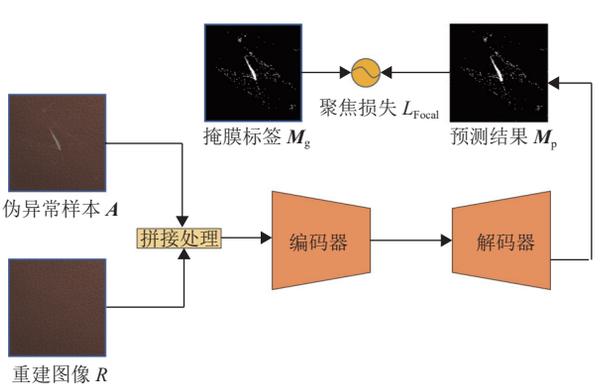


图 3 判别过程的网络结构

Fig.3 The structure of discriminative network

测的异常掩膜结果$\boldsymbol{M}_{\mathrm{p}}$。

对于判别分割网络的损失函数设计，采用了聚焦损失作为约束[16]，具体计算过程见式(11)~(12)。

$$FL(p_n) = -\beta(1-p_n)^{\delta}\log(p_n) \tag{11}$$

$$L_{\mathrm{Focal}} = \frac{1}{N}\sum_{i=1}^{N} FL(p_i) \tag{12}$$

式中：p_n为难易分类样本的概率，p_i表示为第i个像素的输出分类置信度。β为一个调整系数，其作用是平衡正负样本数量，实验中δ则是用于调整对于难易样本的权重，本文在实验中将这两个超参数设置为：$\beta = 0.25$，$\delta = 2$。

2.3　双轨联合的重建特征聚合模块(DRAM)

本文提出了一个双轨联合的重建特征聚合模块DRAM，它旨在对正常样本的特征信息进行过滤，将全局表示连接到每个局部表示来对逐像素关系进行建模，聚合出更有效的瓶颈层特征。如图4所示，所提出的特征聚合模块设计在重建网络的编码器和解码器之间，它包含全局注意力和局部注意力两个分支组成。

受到Hu等[23]提出的SE模块的启发，本文采用全局平均池化层、全连接层、高斯误差线性单元和Sigmoid函数层设计全局注意力。经过编码器降维得到的特征图F首先进入全局平均池化层，并在通道维度对特征图进行特征加权，得到尺寸为$1\times1\times C$的全局信息$\boldsymbol{X}_{\mathrm{g}}$，过程如式(13)所示。

$$\boldsymbol{X}_{\mathrm{g}} = f_{\mathrm{GAP}}(F_{H\times W\times C}) \tag{13}$$

式中：$f_{\mathrm{GAP}}(\cdot)$为全局平均池化操作，计算过程可由式(14)来表述。

$$f_{\mathrm{GAP}}(x) = \frac{1}{(H\times W)}\sum_{i=1}^{W}\sum_{j=1}^{H} x(i,j) \tag{14}$$

式中：H、W、C分别为输入特征图F的高、宽和通道。全局平均池化操作可以提高特征图之间的上下文信息聚合。随后的激励操作中，选用GELU作为激活函数，过程如式(15)所示。

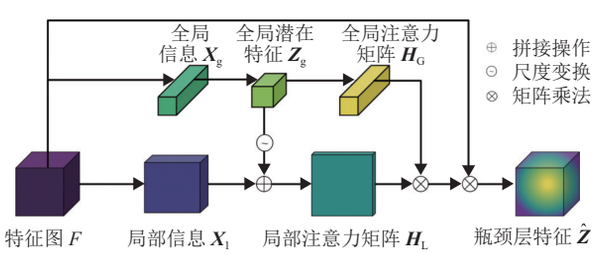


图 4　DRAM的网络架构

Fig.4　Network structure of DRAM

$$\boldsymbol{Z}_{\mathrm{g}} = f_{\mathrm{GELU}}(f_{\mathrm{FC}}(\boldsymbol{X}_{\mathrm{g}})) \tag{15}$$

式中：$f_{\mathrm{FC}}(\cdot)$为全连接卷积层，此处的全连接层可以实现对特征通道数的线性衰减，目的是为了减少GELU函数层中的计算量。因此，经过GELU函数层后，得到输出尺寸为$1\times1\times\gamma\mathrm{C}$的全局潜在特征$\boldsymbol{Z}_{\mathrm{g}}$。$\gamma$是一个控制衰减程度的超参数，实际应用过程中的取值为0.25。

最后，全局潜在特征经过全连接层和Sigmoid函数层后，输出尺度为$1\times1\times C$的全局注意力矩阵$\boldsymbol{H}_{\mathrm{G}}$，过程可由式(16)来表述。

$$\boldsymbol{H}_{\mathrm{G}} = f_{\mathrm{sigmoid}}(f_{\mathrm{FC}}(\boldsymbol{Z}_{\mathrm{g}})) \tag{16}$$

局部注意力分支更多关注于目标及其附近的部分区域，瓶颈层特征首先经过一个通道线性衰减的全连接层和GELU函数层得到尺寸为$H\times W\times\gamma\mathrm{C}$的局部信息$\boldsymbol{X}_{\mathrm{l}}$。全局潜在特征可以帮助模型更加关注样本所在区域，将全局潜在特征$\boldsymbol{Z}_{\mathrm{g}}$与局部信息$\boldsymbol{X}_{\mathrm{l}}$进行聚合，可以避免背景信息带来的干扰，更好地聚焦于样本的细节特征。因此，在局部注意力模块中，将全局注意力模块中的全局信息$\boldsymbol{Z}_{\mathrm{g}}$和局部表示$\boldsymbol{X}_{\mathrm{l}}$进行拼接，得到局部潜在特征$\boldsymbol{Z}_{\mathrm{l}}$，整个过程可由式(17) 来表述。

$$\boldsymbol{Z}_{\mathrm{l}} = \mathrm{Concat}(\boldsymbol{X}_{\mathrm{l}}, f_{\mathrm{reshape}}(\boldsymbol{Z}_{\mathrm{g}})) \tag{17}$$

式中：$f_{\mathrm{reshape}}(\cdot)$为维度变换操作，旨在将全局潜在特征的尺寸从$1\times1\times\gamma\mathrm{C}$扩展成$H\times W\times\gamma\mathrm{C}$，Concat则是在通道维度将两类特征进行拼接。随后依次输入到全连接层和Sigmoid函数层，得到尺寸为$H\times W\times1$的局部注意力矩阵$\boldsymbol{H}_{\mathrm{L}}$。整个过程如式(18) 所示。

$$\boldsymbol{H}_{\mathrm{L}} = f_{\mathrm{sigmoid}}(f_{\mathrm{FC}}(\boldsymbol{Z}_{\mathrm{l}})) \tag{18}$$

结合上述双轨分支中得到的全局注意力矩阵和局部注意力矩阵，潜在空间中的瓶颈层特征可以表述为

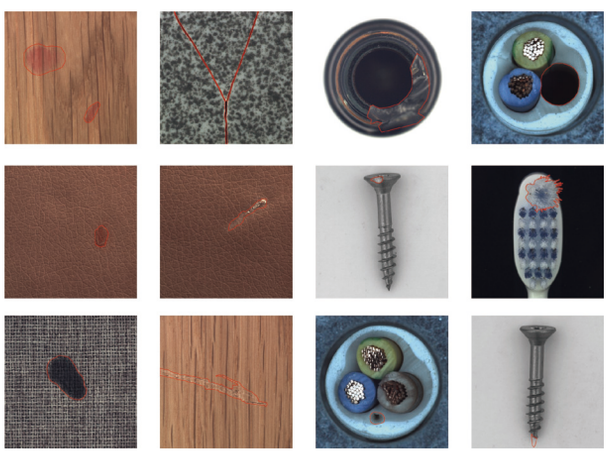
$$\hat{\boldsymbol{Z}} = F\times\boldsymbol{H}_{\mathrm{G}}\times\boldsymbol{H}_{\mathrm{L}} \tag{19}$$

3　实验结果及讨论

3.1　数据集

MVTec数据集[24]是一个关注工业场景下的异常检测数据集。它包含了15种不同对象和纹理类别，如瓶子、螺丝钉、皮革、地毯等。在每个类别中，训练集仅包含了无缺陷的图像，而测试集中不仅有无缺陷的图像，还包含73种不同的场景下的缺陷类别，涵盖了表面缺陷、结构缺陷等真实工业场景下比较关注

的异常实例，可以更加全面地模拟真实工业场景的异常检测。为了验证模型在实际工业场景下的性能，如图5所示，本文实验中选用了地毯、皮革、瓷砖和木材作为表面缺陷的异常样本，如颜色污染、表面划痕等；结构缺陷方面选用了瓶子、牙刷、电缆和螺丝钉，如瓶口破损、线缆缺失等。



(a) 颜色污染 (b) 划痕裂缝 (c) 破损缺陷 (d) 结构缺失

图5 MVTec数据集的部分缺陷

Fig.5 Defects in the MVTec dataset

3.2 实验细节

首先，采用2.1节提出的异常样本生成模块SDGM对训练数据进行预处理。整个过程可以分成两个阶段，第一阶段通过随机过程产生各种形状的掩膜区域，第二阶段则是根据缺陷语义的引导在掩膜区域中合成具体的缺陷。本文实验中针对每种类别均提前设置了4种缺陷语义，确保异常样本生成的多样性，从而提高模型的泛化能力。其次，采用经过预处理的数据训练两阶段异常检测模型，重建网络和判别分割网络是同时训练的，但是两者的梯度计算是独立的。模型采用Adam优化器进行训练，训练开始时设置学习率为10^{-4}，当训练周期到达400个迭代周期和600个迭代周期时，学习率分别衰减为10^{-5}和10^{-6}，直到第700个迭代周期时，结束训练。本文所有实验均使用Pytorch1.13框架，代码编写采用Python3.7，模型训练过程使用的种子数为42，图像尺寸为256，批大小为8。推理验证的CUDA版本为11.6，GPU为RTX3090。

3.3 评价指标

异常检测任务通常使用受试者工作特性曲线下面积(Area Under the Receiver Operating Characteristic curve, AUROC)来评估模型的效果。AUROC通过接受者操作特征曲线与坐标轴之间的面积大小来评价模型的性能。AUROC的数值越大，模型检测异常区域的正确率越高。在异常检测任务中，图像级别的AUROC结果是评价模型判断异常的能力，像素级别的AUROC结果则可以体现模型分割出具体异常缺陷区域的能力。此外，为了进一步反映模型在分割精度和误检几率上的表现，可以采用像素级别的F1-score结果来衡量。本文评估阶段用f_1表示最终的像素级别的F1-score，其计算过程如式(20)所示。

$$f_1 = \frac{1}{N}\sum_{k=1}^{N}\frac{2\times P_k \times R_k}{P_k + R_k} \tag{20}$$

式中：P_k和R_k分别为第k个样本的精准率和召回率，可由式(21)和(22)来表述。

$$P_k = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{21}$$

$$R_k = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{22}$$

式中：TP为缺陷像素被正确预测的数量，FP为缺陷像素被错误预测的数量，FN为非缺陷像素被错误预测的数量。同时，F1-score还是异常检测中作为衡量模型鲁棒性的重要指标，它可以通过精度与召回率的平衡反映出模型的误检与漏检的情况，具体来说，F1-score数值越接近1表示模型的鲁棒性越好。

因此，本文采用图像级别的AUROC值和像素级别的AUROC值作为实验评估指标，图像级别的AUROC反映模型判别样本检测异常的准确性，像素级别的AUROC反映模型对异常区域分割的精确度。最后，引入了像素级别的F1-score作为模型鲁棒性的评价。

3.4 实验结果与分析

3.4.1 对比实验

本文将DADNet和4个异常检测模型FastFlow、PaDim、RD和STFPM进行性能比较，为确保对比实验的公平性，每个模型进行3次评估实验，最终结果取34次实验的平均值。

表1呈现了在MVTec数据上本文提出的模型和4个异常检测模型的图像级别的AUROC性能评估结果，其中各类别最优性能得分以加粗字体标出。相比于其他异常检测模型相比，DADNet在8类别的平均图像AUROC取得最高得分，有4个类别的图像AUROC分数排名第一。这4个类别中主要集中在表面缺陷的代表类别，说明DADNet对于表面缺陷的异常检测能力较为突出。

值得注意的是，在Screw类别中，其他4个检测模型的分数表现都出现明显的波动，而DADNet比第2名的Padim模型高出14.1个百分点，领先优势非常明

表 1　在MVTec数据集上的图像AUROC评估结果
Table 1　AUROC-Image evaluation on the MVTec dataset　%

类别	FastFlow	PaDim	RD	STFPM	DADNet
Carpet	97.9	**99.8**	99.6	95.4	95.8
Leather	**100.0**	**100.0**	87.8	98.9	**100.0**
Tile	96.7	98.1	83.3	94.9	**99.1**
Wood	97.8	99.2	98.9	96.1	**99.8**
Bottle	**100.0**	99.9	99.8	97.9	99.4
Toothbrush	84.4	96.1	96.7	**99.7**	99.1
Cable	89.1	92.7	**98.2**	83.8	94.4
Screw	52.1	85.8	77.8	83.5	**99.9**
Average	89.8	96.5	92.8	93.8	**98.4**

显。在所有类别的图像AUROC结果中，DADNet始终保持94%以上的得分，这也侧面证明DADNet具有更稳定的检测性能。

表2展示了各模型在MVTec数据集上的像素AUROC评估结果，该指标反映的是模型对于缺陷区域的分割精度，其中各类别最优性能得分以加粗字体标出。从整体上看，DADNet在所有类别的平均像素AUROC上再次取得最高得分，说明DADNet具备稳定的分割性能。在Tile类别上，DADNet相比其他4个异常检测模型有着更好的表现，领先第2名的STFPM模型3.4个百分点。

为了更全面展现各个模型的综合性能，表3汇总了各模型在MVTec数据集上的像素F1-score。像素F1-score可以一定程度反映出误检的情况，得分越高说明模型的稳定性和泛化性越好。可以看到，DADNet在所有类别上的平均像素F1-score表现远超过其他4个异常检测模型，并且有7个类别的像素F1-score排在第一。尤其在Screw类别上，其他4个模型的像素F1-score均不超过50%。结合表1的图像AUROC结果来看，可以推断出这是因为模型对于该类别的缺陷检测较差，存在大量的漏检现象。

3.4.2　可视化结果分析

为了进一步分析DADNet取得优势的原因，图6展示了MVTec测试数据集中的部分缺陷异常分割结果，包含了颜色污染、表面划痕、结构缺失等缺陷。其中，GroundTruth表示缺陷区域的掩膜标签，与掩膜标签相比，DADNet表现出最好的分割结果。一方面，对于结构缺陷，DADNet能在减少误检的情况下更加精细地分割，这印证了DADNet具有更好的表征瓶颈层特征，可以更精确地修复细节信息。反观其他4个异常检测模型，它们在结构缺陷的分割中都存在过多的误检情况，尤其是对于图6(a)中第4行的螺丝钉(Screw类)，PaDim模型甚至无法检测出异常缺陷，由此可知，造成3.4.1节中AUROC-Pixel领先的原因是DADNet在训练过程中具有更高效的监督信息，使用SDGM模块可以实现更加真实的缺陷样本模拟，进一步提升了模型的判别能力。

另一方面，对于表面缺陷，DADNet的分割效果依旧稳定，如图6(b)中第2行的皮革和第3行的瓷砖，其他4个模型先是在皮革的检测中出现漏检区域，其次在瓷砖的检测中存在大量的误检，而DADNet在这两个类别的测试中都展现出最佳的视觉分割结果。因此，与其他模型相比，DADNet的异常检测能力更加稳定，对于缺陷的分割更加精细，在不同工业场景下的应用也更具潜力。

表 2　在MVTec数据集上的像素AUROC评估结果
Table 2　AUROC-Pixel evaluation on the MVTec dataset　%

类别	FastFlow	PaDim	RD	STFPM	DADNet
Carpet	98.3	**99.0**	98.9	98.6	97.5
Leather	**99.6**	99.0	98.4	99.1	99.3
Tile	94.4	94.1	86.7	94.6	**98.3**
Wood	95.6	94.1	93.9	94.9	**96.7**
Bottle	98.3	98.2	98.1	97.1	**99.1**
Toothbrush	97.9	98.8	**99.0**	98.3	98.7
Cable	95.4	**96.7**	96.5	89.8	95.8
Screw	92.6	98.4	98.7	98.3	**99.6**
Average	96.5	97.3	96.3	96.3	**98.1**

表 3　在MVTec数据集上的像素F1-score评估结果
Table 3　F1-score-Pixel evaluation on the MVTec dataset　%

类别	FastFlow	PaDim	RD	STFPM	DADNet
Carpet	57.3	58.1	59.6	**64.2**	62.2
Leather	56.0	49.5	45.4	55.8	**65.7**
Tile	56.9	54.0	63.2	70.4	**87.8**
Wood	55.7	47.2	52.2	59.2	**65.7**
Bottle	67.0	72.2	71.0	78.5	**83.5**
Toothbrush	47.9	57.3	56.6	**60.9**	59.7
Cable	54.7	45.8	52.6	61.5	**67.3**
Screw	6.1	22.4	45.6	39.0	**69.3**
Average	50.2	50.8	55.8	61.2	**70.2**

另一方面，对于表面缺陷，DADNet的分割效果依旧稳定，如图6(b)中第2行的皮革和第3行的瓷砖，其他4个模型先是在皮革的检测中出现漏检区域，其次在瓷砖的检测中存在大量的误检，而DADNet在这两个类别的测试中都展现出最佳的视觉分割结果。因此，与其他模型相比，DADNet的异常检测能力更加稳定，对于缺陷的分割更加精细，在不同工业场景下的应用也更具潜力。

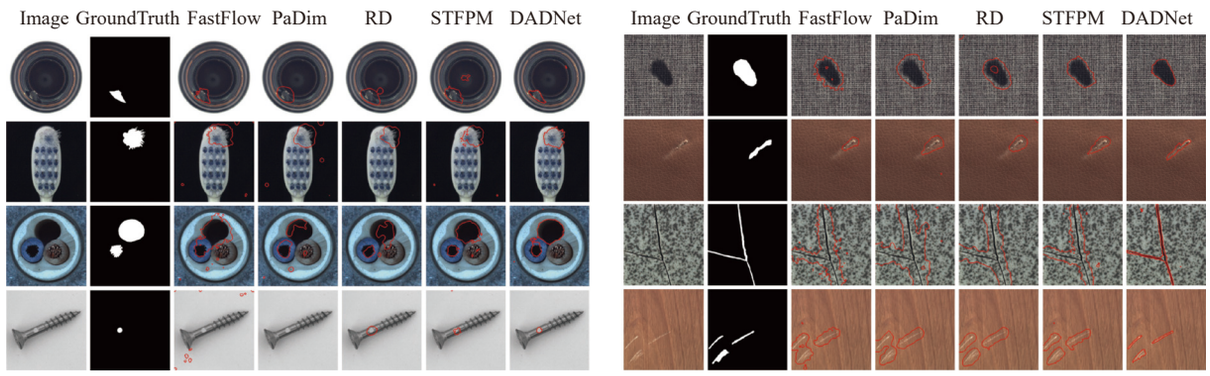
3.4.3　消融实验

本文主要针对SDGM和DRAM两个模块设计如下消融实验：

(1) 在保持基线主干网络结构不变的前提，比较通过纹理或者噪声合成伪异常样本的方法与SDGM的性能差异；(2) 在保持基线主干网络结构不变的前提，比较加入DRAM前后的网络性能提升。

本文消融实验的结果如表4所示，为了比较模型的总体性能，这里的结果仅对所有类别上的平均结果进行分析。其中，Baseline是以U-Net为主干网络，并包含双编码器–解码器模块的异常检测网络。另外，通过纹理或者噪声生成缺陷样本的模块表示为DGM(Defect Generation Module)。

从表4中可以发现，SDGM相比于以往的伪异常



(a) 结构缺陷分割结果 (b) 表面缺陷分割结果

图 6 测试集上的可视化结果

Fig.6 Visualization results on the test set

表 4 消融实验结果

Table 4 Ablation experimental results %

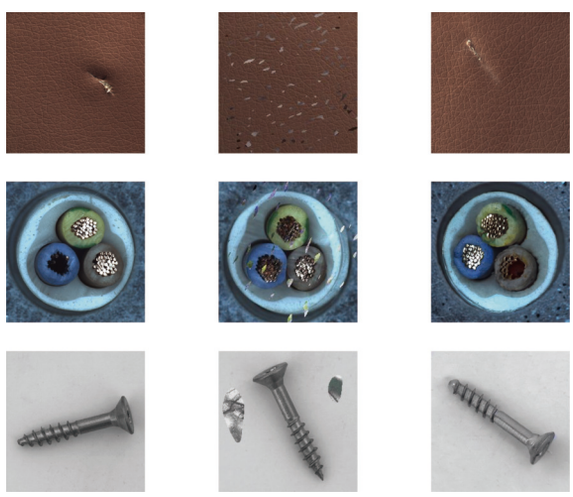
实验	AUROC-Image	AUROC-Pixel	F1-score-Pixel
Baseline+DGM	78.0	79.3	25.4
Baseline+SDGM	97.2	97.4	60.9
Baseline+DGM+DRAM	97.6	97.4	67.9
DADNet(Ours)	**98.4**	**98.3**	**70.2**

样本生成方法，取得更好的图像AUROC和像素AUROC，并且像素F1-score的结果得到大幅提升。结合图7所示的伪异常样本对比图可知，DGM合成的缺陷是通过添加噪声实现的，与真实缺陷图之间存在巨大差异。而SDGM可以合成与工件相关的语义缺陷，如划痕、线缆缺失和破损等，这些伪异常样本在视觉观感上更接近真实缺陷图。这也进一步说明融合语义信息的伪异常样本可以提供足够的对抗信息，帮助模型更好地约束正常样本的分布边界。为了验证所提出的DRAM模块的有效性，在Baseline和DGM的基础上，加入DRAM，同样可以看到像素F1-score结果提升了42.5个百分点。而在AUROC的分数上，DRAM模块的加入，也为模型带来了和SDGM模块嵌入后相近的分数结果。

值得注意的是，在像素F1-score指标上，DRAM模块的表现要比SDGM模块要好，这说明加入DRAM后的模型可以更加稳定。由于SDGM生成的异常样本相比于传统伪异常样本方法更加逼近正常样本，而Baseline模型中缺少约束潜在空间的模块，直接在Baseline上利用SDGM生成的异常样本做训练也会遇到模型泛化能力的瓶颈。因此，表4的DADNet结果则充分证明，经过DRAM聚合特征，潜在空间中正常特征与异常特征混淆的问题得到缓解，模型的泛化能力进一步提高。

4 结语

本文面向工业场景提出了一个基于扩散生成的



(a) 真实缺陷 (b) DGM 合成的噪声缺陷 (c) SDGM 合成的缺陷

图 7 伪异常样本对比图

Fig.7 Comparison of the pseudo anomaly samples

两阶段异常检测网络DADNet，它沿用图像修复的框架，分为异常修复和异常检测两部分，并且该网络适用于不同类别的工件缺陷异常检测。针对以往伪异常样本缺乏缺陷语义信息的问题，本文提出了语义引导的缺陷生成模块SDGM，为异常修复的训练过程提供更具对抗性的异常样本，从而更好地约束正常样本的最小化边界。同时，DADNet针对潜在空间中正常特征与异常特征的混淆问题，提出DRAM模块，通过全局表示和局部表示的聚合进行建模，以提升重建模型的鲁棒性。基于MVTec数据集的对比实验与消融实验证明了DADNet的优异性能及其网络中各模块的有效性。

参考文献：

[1] 李忠海, 白秋阳, 王富明, 等. 基于语义分割的钢轨表面缺陷实时检测系统[J]. 计算机工程与应用, 2021, 57(12): 248-256.
LI Z H, BAI Q Y, WANG F M, *et al*. Real-time detection

system of rail surface defects based on semantic segmentation[J]. Computer Engineering and Applications, 2021, 57(12): 248-256.

[2] ZAVRTANIK V, KRISTAN M, SKOČAJ D. DRAEM—A discriminatively trained reconstruction embedding for surface anomaly detection[C] //2021 IEEE/CVF International Conference on Computer Vision (ICCV). Online: IEEE Computer Society, 2021: 8310-8319.

[3] YOUKACHEN S, RUCHANURUCKS M, PHATRAPOMNANT T, *et al*. Defect segmentation of hot-rolled steel strip surface by using convolutional auto-encoder and conventional image processing[C]//2019 10th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES) . New York: IEEE, 2019: 1-5.

[4] KANG G, GAO S, YU L, *et al*. Deep architecture for high-speed railway insulator surface defect detection: denoising autoencoder with multitask learning[J]. IEEE Transactions on Instrumentation and Measurement, 2018, 68(8): 2679-2690.

[5] 黄剑航, 王振友. 基于特征融合的深度学习目标检测算法研究[J]. 广东工业大学学报, 2021, 38(4): 52-58.
HUANG J H, WANG Z Y. A research on deep learning object detection algorithm based on feature fusion[J]. Journal of Guangdong University of Technology, 2021, 38(4): 52-58.

[6] CHOW J K, SU Z, WU J, *et al*. Anomaly detection of defects on concrete structures with the convolutional autoencoder[J]. Advanced Engineering Informatics, 2020, 45: 101105.

[7] NAKAZAWA T, KULKARNI D V. Anomaly detection and segmentation for wafer defect patterns using deep convolutional encoder–decoder neural network architectures in semi-conductor manufacturing[J]. IEEE Transactions on Semi-conductor Manufacturing, 2019, 32(2): 250-256.

[8] BALZATEGUI J, ECIOLAZA L, MAESTRO-WATSON D. Anomaly detection and automatic labeling for solar cell quality inspection based on generative adversarial network[J]. Sensors, 2021, 21(13): 4361.

[9] TAYEH T, ABURAKHIA S, MYERS R. Distance-based anomaly detection for industrial surfaces using triplet networks[C]//2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). Online: IEEE, 2020: 0372-0377.

[10] ZAVRTANIK V, KRISTAN M, SKOČAJ D. Reconstruction by inpainting for visual anomaly detection[J]. Pattern Recognition, 2021, 112: 107706.

[11] SCHLÜTER H M, TAN J, HOU B, *et al*. Natural synthetic anomalies for self-supervised anomaly detection and localization[C]//European Conference on Computer Vision. Cham: Springer, 2022: 474-489.

[12] ZHAO Z, LI B, DONG R, *et al*. A surface defect detection method based on positive samples[C]//PRICAI 2018: Trends in Artificial Intelligence: 15th Pacific Rim International Conference on Artificial Intelligence. Switzerland: Springer International Publishing, 2018: 473-481.

[13] ZHANG D, GAO S, YU L, *et al*. DefGAN: defect detection GANs with latent space pitting for high-speed railway insulator[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-10.

[14] YU J, KIM D Y, Lee Y, *et al*. Unsupervised pixel-level road defect detection via adversarial image-to-frequency transform[C]//2020 IEEE Intelligent Vehicles Symposium (IV) . Stuttgart: IEEE, 2020: 1708-1713.

[15] YU J, ZHENG Y, WANG X, *et al*. Fastflow: unsupervised anomaly detection and localization via 2d normalizing flows[EB/OL]. arXiv: 2111.07677(2021-11-16) [2024-05-30]. https://arxiv.org/abs/2111.07677.

[16] DEFARD T, SETKOV A, LOESCH A, *et al*. PaDiM: a patch distribution modeling framework for anomaly detection and localization[C] //International Conference on Pattern Recognition. Online: Springer: 2021: 475-489.

[17] DENG H, LI X. Anomaly detection via reverse distillation from one-class embedding[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) . New Orleans: IEEE, 2022: 9737-9746.

[18] WANG G, HAN S, DING E, *et al*. Student-teacher feature pyramid matching for unsupervised anomaly detection[EB/OL]. arXiv: 2103.04257(2021-10-28) [2023-12-16]. https://arxiv.org/abs/2103.04257

[19] RADFORD A, KIM J, HALLACY C, *et al*. Learning transferable visual models from natural language supervision [C]//International Conference on Machine Learning. Online: PMLR: 2021: 8748-8763.

[20] RAUNIYAR A, RAJ A, KUMAR A, *et al*. Text to image generator with latent diffusion models[C]//2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN) . Ghaziabad: IEEE, 2023: 144-148.

[21] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]//Neural Information Processing Systems. Long Beach: Curran Associates, 2017: 5998-6008.

[22] WANG Z, BOVIK A C, SHEIKH H R, *et al*. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[23] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 7132-7141.

[24] BERGMANN P, FAUSER M, SATTLEGGER D, *et al*. MVTec AD — A comprehensive real-world dataset for unsupervised anomaly detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) . Long Beach: IEEE, 2019: 9592-9600.

（责任编辑：杨耀辉　英文审核：费伦科）