Abstract

一句话点题+介绍网络结构

自动缺陷检测技术得到了极大地提升，随着深度学习的逐步渗透。然而很多领域中缺陷检测任务依然有挑战性，由于其产品的多样性。在这篇文章中，我们关注于轮胎工业中的缺陷检测并提出了一个...网络，（which）它包含两个并行的子网络（network或branch）用来感知（挖掘）多尺度的缺陷特征。具体的（Specifically），蕴藏在深网络层中的（需要深层网络获取的）缺陷形状和位置信息被挖掘通过结构感知网络，（which）以传统的（已存在的）全卷积网络为框架。与此同时，为了弥补深层网络造成的细节丢失，一个新奇的纹理感知网络被用来（is developed for）尽可能的挖掘cover小尺寸的缺陷。之后，通过融合含有结构和细节信息的特征得到最后的密集预测结果。实验结果表明，我们的网络可以得到令人满意的效果，特别是小缺陷的检测。

Though automatic detection method has been tremendous improved, with the gradual penetration of deep learning. Defect detection in many industrial processes is one of the remaining challenging tasks due to the diversity of its products. In this work, we focus on detection tasks in the tire industry and develop a *Multi-scale Defect Detection Network (MDDN)*, which contains two parallel sub-networks to capture multi-scale defect features. Specifically, high-abstracted semantic features containing defect shapes and locations are mined via a *Semantic-aware Sub-network*, simplified by an off-the-shelf fully convolutional network. Furthermore, to complement the details filtered by the deep network, a novel *Texture-aware Sub-network* is used to exploit the small size of the cover edge features and small defects as much as possible. Finally, the pixel-wised detection results are obtained by fusing features with semantic and texture information. Extensive experiments demonstrate that *MDDN* can produce comparable results and achieve significantly performance improvement in small defects detection.

不像之前的网络，它可以同时定位并分割缺陷。在...数据集上的实验结果表明...

Introduction

自动检测技术，被用来提高质量和加速生产，已经成为不可或缺的一部分在许多工业领域，例如\*\*\*，\*\*\*。特别是在轮胎产品中，缺陷检测任务一直受到广泛关注并且已经有很多研究被提出[\*\*\*]（在过去的几十年中）。most real-world 检测通常首先从获得的X-ray 图像中挖掘出缺陷区域，which（由\*\*\*造成的），然后有缺陷的产品根据缺陷的位置、形状等信息被分等级处理。由于已存在方法的局限性和图像的特殊性[\*，\*]，大多依然采用人工的形式，造成了\*\*（人工的缺点）。因此轮胎缺陷检测依然是最有挑战的工业检测任务之一。

Automatic defect detection, used to improve quality and accelerate production, has become an indispensable part in industrial processes, such as fabrics[1,2,3], steel[4], semiconductors[5], and solar wafers[6]. Especially in tire manufacturing, numerous detection algorithms have been proposed [7,8,9,10,11] and aroused extensive attention over the past two decades. In most real-world applications, tire defect detection is first carried out by deriving the defective region from tire X-ray images, which contains various types of defects caused by unclean raw materials and undesired manufacturing facilities[12]. Then, the defective product is hierarchical processed according to the location and size of defects. Due to unique properties of the tire image, for instance complexity and low-quality, illustrated in previous study[13, 14], most inspection processes are performed by human observers, which increases the risk and reduces the efficiency. Therefore, tire defect detection remains one of the most challenging inspection tasks.

目前存在基于计算机视觉的方法大多致力于通过区分缺陷区域与正常区域检测缺陷，一个关键的问题是。。。a key issue for such methods is texture feature extraction. 从特征表示的角度，这些方法可以被大体分为三类：固定特征，固定变换核学习系数，学习特征。\*\*\*\*；\*\*\*\*；\*\*\*\*；然而，大量实验结果表明，这种方法不够精确，对边缘细节和小缺陷不敏感。这个问题同样出现在使用深度网络分割自然图像的任务中。（这与深度网络分割自然图像时所遇到的问题一致。）

At present, existing computer vision based detection methods are mostly devoted to distinguish difference between defective regions and background (defective-free regions). Hence a key issue for such methods is feature extraction. Guo et al. [15] exploited a local kernel regression descriptor to derive feature vectors. By comparing the dissimilarity of the corresponding feature between one pixel and its neighbors, anomaly pixels can be located and segmented, even in the tread image. Nevertheless, this method is not suitable for real-time detection tasks because of the high computational complexity. A component decomposition based method was proposed in [15] , which separated the background from the image by means of two designed filters. Then through an adaptive thresholding processing, defects were derived from the residual image. Besides, Independent component analysis(ICA) was also used for defect detection tasks[8,9]. A major disadvantage of these fundamental methods is the limitation of the information contained in low-level clues and domain features. ~~A major disadvantage of these fundamental~~ ~~methods that rely on low-level clues is the limitation of representation capabilities.~~ To address the limitation, Zhang et al. [] and Zhang et al. [][] introduced radon transform and mulit-scale transform, for instance curvelet and wavelet transform, in detection tasks respectively. Furthermore, optimized edge detection and total variation algorithm are used to achieve more accurate results[]. Zhao et al.[] proposed a multiple kernel learning method, which combined various transform kernels to get more differentiated information. However, the representation capability of fixed kernels is not comprehensive enough. In addition, transform process is computationally expensive. Recently, Cui et al. [] attempted to classify tire defects by means of convolutional neural networks(CNN), which has outstanding performance in the recognition and segmentation tasks of natural images. With the excellent feature extraction capability of deep network, Wang et al. [] further implemented the detection and segmentation in tire images by a fully convolutional network(FCN). However, FCN is not sensitive to small defects and edge details, which is similar to that in dealing with natural image tasks.

为了解决这个问题，已经有很多多尺度的方法被提出在公共数据集中。（成功的开始介绍之前想过的多尺度方法）。这些方法虽然在一定程度上缓解了小目标的检测，但是两个关键的问题：1，没有考虑过比例的问题。2，细节纹理的提取只在浅层，而浅层过多会导致参数爆炸，过少又提取不到关键特征。

To overcome the shortcomings of insensitivity to small defects and detailed textures, many methods have been proposed in benchmark datasets. Most of them are based on multi-scale strategies and can be roughly classifified into image pyramids and in-network feature hierarchies. Image pyramids illustrated in Figure 1a[] were extensively used in the era of hand-crafted features[]. Even if the crafted features have largely been replaced by self-learning features, multi-scale testings on the image pyramid are still used to verify the adaptability and robustness. Nevertheless, image pyramid based methods is impractical for real applications due to the considerable increase in inference time. In-network feature hierarchies are formed by the forward propagation within deep convolutional networks(ConvNets). Due to a series of sub-sampling layers, in-network hierarchies produce feature maps of different spatial resolutions, with an multi-scale and pyramid shape[]. By fusing these multi-scale feature maps, the texture feature in shallow layers and the semantic information contained in deep layers can be perceived. The Single Shot Detector (SSD) [] is one of the first attempts at combining predictions from these features maps to detect objects of various sizes. Generally, shallow features are used to predict small objects, and deep features with large receptive fields are used detect large objects. However, the lack of semantic information is harmful to the detection of small targets in shallow layers. Another fusing way can effectively address this problem by concatenating multi-scale features and detecting on top of the expanded feature maps, as shown by the red line in Figure 1b[]. For example, FCN-8s and FNC-16s defined a skip architecture to produce more accurate segmentation. Similar top-down skip architectures are popular in recent research[]. There exists a basic problem that it is still not enough to mine the detail texture in these structures[]. Bai et al. proposed a novel multi-task generative adversarial network (MTGAN), which improve the detection performance by up-sampling a small object to a larger scale using super-resolution.

Inspired by MTGAN, we construct a end-to-end network named Multi-scale Defect Detection Network (MDDN) consisting of a semantic-aware sub-network and a texture-aware sub-network. Based on tire X-ray images, a image patch strategy is adopted in the texture-aware sub-network. Unlike natural image patches, defects (objects) are still significant and discernible in the tire image patches, as shown in Figure 1(c). On the one hand, The proportion of the area of defective regions in images increases, which is advantageous for better capturing of detailed information. As shown in Figure 1(b), for a 256\*256 tire image, the proportion of defects is increased from \*\*\* to \*\*\*. On the other hand, image patches as input data can be used without reducing the number of parameters in the case of reducing the pooling layer. On the other hand, sub-sampling layers can be discarded to retain more shallow features without increasing the parameters.

8-19日改

To overcome these shortcomings, many methods have been proposed in benchmark datasets. Most of them are based on multi-scale strategies and can be roughly classifified into image pyramids and in-network feature hierarchies. Image pyramids were directly scaled to get multi-scale images and extensively used in the era of hand-crafted features \cite{lowe2004distinctive,dalal2005histograms}, as shown in Figure 1(a). Even if the crafted features have largely been replaced by self-learning features, multi-scale testings on the image pyramid are still used to verify the adaptability and robustness ({\it e.g.}, \cite{he2016deep}). Nevertheless, image pyramid based methods is impractical for real applications due to the considerable increase in inference time. In-network feature hierarchies are formed by the forward propagation within deep ConvNets. Through several of sub-sampling layers, in-network hierarchies produce feature maps of different spatial resolutions, with an multi-scale and pyramid shape\cite{lin2017feature}. By fusing these multi-scale feature maps, features in shallow and deep layers can be perceived. The Single Shot Detector (SSD) \cite{liu2016ssd} is one of the first attempts at combining predictions from these features maps to detect objects of various sizes. Generally, shallow features are used to predict small objects, and deep features with large receptive fields are used to detect large objects. However, the lack of semantic information is harmful to the detection of small targets in shallow layers. Another fusing way can effectively address this problem by concatenating multi-scale features and detecting on top of the expanded feature maps, as shown by the red line in Figure 1b[]. For example, FCN defined a skip architecture to produce more accurate segmentation. Similar top-down skip architectures are popular in recent research\cite{newell2016stacked,ghiasi2016laplacian}. There exists a basic problem that it is still not enough to mine the detail texture in these structures\cite{zhou2018scale}. Bai {\it et al}. proposed a novel multi-task generative adversarial network (MTGAN), which improve the detection performance by up-sampling a small object to a larger scale using super-resolution.

Inspired by MTGAN, we construct a end-to-end network named Multi-scale Defect Detection Network (MDDN) consisting of a semantic-aware sub-network and a texture-aware sub-network.

To detect small tire defects, image patches \cite{bai2018sod,chen2016attention} are fed in the texture-aware sub-network. Unlike natural image patches, defects (objects) are still significant and discernible in the tire image patches, illustrated in Figure 1(d). In addition, as shown in Figure 1(c), the proportion of defects in a 256 $\times$ 256 tire image increases from 124/65535 to 124/4096, which is advantageous for better capturing of detailed information.

改：

To overcome these shortcomings, many methods have been proposed in benchmark datasets. Most of them are based on multi-scale strategies and can be roughly classifified into image pyramids and in-network feature hierarchies. Image pyramids were directly scaled to get multi-scale images and extensively used in the era of hand-crafted features \cite{lowe2004distinctive,dalal2005histograms}, as shown in Figure 1(a). With the popularity of self-learning feature representation, image pyramid based methods is impractical for real applications due to the considerable increase in inference time. In-network feature hierarchies are formed by the forward propagation within deep ConvNets. Through several of sub-sampling layers, in-network hierarchies produce feature maps of different spatial resolutions, with an multi-scale and pyramid shape\cite{lin2017feature}. The Single Shot Detector (SSD) \cite{liu2016ssd} is one of the first attempts at combining predictions from these features maps to detect objects of various sizes. Generally, shallow features are used to predict small objects, and deep features with large receptive fields are used to detect large objects. However, the lack of semantic information is harmful to the detection of small targets in shallow layers. Another fusing way can effectively address this problem by concatenating multi-scale features and detecting on top of the expanded feature maps, as shown by the red line in Figure 1b[]. For example, FCN defined a skip architecture to produce more accurate segmentation. Similar top-down skip architectures are popular in recent research\cite{newell2016stacked,ghiasi2016laplacian}. Since the training goal of the entire network is to extract abstract semantics, there exists a basic problem that it is still not enough to mine the detail texture in these structures\cite{zhou2018scale}. CrowdNet \cite{zhou2018scale} combines shallow and deep features derived from two different structures to overcome this shortcoming. The capability to extract detail features is further enhanced by a parallel shallow subnetwork.

Inspired by CrowdNet, we construct a end-to-end network named Multi-scale Defect Detection Network (MDDN) consisting of a semantic-aware sub-network and a texture-aware sub-network.

图1：

1. shows the image pyramid where the image is scaled to different size. (b) indicates in-network feature hierarchies. Prediction results can be derived from each layer. (c) represents multi-size tire image patches, which can increase the relative scale of small defects and backgrounds.

Introduction v4

第一段：1. 检测很重要，2.在很多领域都有研究，3. 轮胎也有很多研究，4.

Introduction v3

1. 一句话介绍，自动检测的重要性，\*\*\*，\*\*\*，\*\*\*领域均有研究，其中轮胎检测一直广泛受到关注，尽管\*\*\*\*（把一些不好融合到related work中的轮胎检测一并说了）都做出了大量的贡献（就大概这个意思吧），但依然存在很大的挑战。

在实际的工业应用中的轮胎检测环节，根据X光的射线图像（如图所示）定位并检测缺陷（类型、形状、位置、面积等用来解释挑战），之后根据缺陷的位置和面积决定对轮胎进行对应的处理。多数工厂普遍采用人工检测，效率低，不安全，浪费资源......

1. 为了解决这个问题（任务或现状），近几年有一些研究被提出和使用，可以分为\*\*\*和\*\*\*和\*\*\*。分别简单介绍一下。（包含机器学习的方法，只介绍机器学习在轮胎上的应用即可，其他的不用说）。\*\*\*证明了深度学习（分割方法）在轮胎检测的可行性，但是结果表明这种pixel-to-pixel的网络对细节信息和小目标检测效果差，这同样也是自然图像上遇到的问题（和自然图像中的表现一致）（这里重新组织一下，重点是要把话题转移到自然图像、多尺度上）。
2. 为了解决这个问题，已经有很多多尺度的方法被提出在公共数据集中。（成功的开始介绍之前想过的多尺度方法）。这些方法虽然在一定程度上缓解了小目标的检测，但是忽略了一个本质的原因...\*\*\*首先引入超像素的方法解决小缺陷的问题，为占比问题提供了一个新的思路并证实了占比的重要性。\*\*\*在小人脸检测任务中采用\*\*\*策略，在原有的金字塔模型上固定目标框，缩放图片（这样做主要是因为人脸检测中的图像允许他这样做，通过这一条再说对其他图像并不适用）。针对轮胎图像的特点（可以简单的列举一下特点，或者直接引用某篇文献），我们提出了一种......方法并新建了一个网络，经常被用在传统方法的预处理中，\*\*\*、\*\*\*、\*\*\*也使用了这种方法。（大概列几个文献，证明可行性）。

Introduction v2

1. 一句话，自动检测的重要性，\*\*\*，\*\*\*，\*\*\*领域均有研究，轮胎检测依然持续受到关注，其中\*\*\*\*（把一些不好融合到related work中的轮胎检测一并说了）都做出了大量的贡献（就大概这个意思吧）。
2. 在实际的工业应用中，轮胎检测的基本流程：a.拿到x光射线的图像（如图所示）；b.检测；c.根据检测结果决定对轮胎进行怎么样的处理（写流程太傻了，中心思想是“需要根据检测结果对轮胎进行下一步处理”，突出检测到缺陷大小的重要性）。人工检测，效率低，不安全，浪费资源......
3. 目前已存在的检测算法大体可以分为\*\*\*和\*\*\*和\*\*\*。分别简单介绍一下。（包含机器学习的方法，只介绍机器学习在轮胎上的应用即可，其他的不用说）。\*\*\*证明了深度学习（分割方法）在轮胎检测的可行性，但是结果表明这种pixel-to-pixel的网络对细节信息和小目标检测效果差，这同样也是自然图像上遇到的问题（和自然图像中的表现一致）（这里重新组织一下，重点是要把话题转移到自然图像、多尺度上）。
4. 为了解决这个问题，已经有很多多尺度的方法被提出在公共数据集中。（成功的开始介绍之前想过的多尺度方法）。这些方法虽然在一定程度上缓解了小目标的检测，但是忽略了一个本质的原因...\*\*\*首先引入超像素的方法解决小缺陷的问题，为占比问题提供了一个新的思路并证实了占比的重要性。\*\*\*在小人脸检测任务中采用\*\*\*策略，在原有的金字塔模型上固定目标框，缩放图片（这样做主要是因为人脸检测中的图像允许他这样做，通过这一条再说对其他图像并不适用）。针对轮胎图像的特点（可以简单的列举一下特点，或者直接引用某篇文献），我们提出了一种......方法并新建了一个网络，经常被用在传统方法的预处理中，\*\*\*、\*\*\*、\*\*\*也使用了这种方法。（大概列几个文献，证明可行性）。

Introduction v1

（1）轮胎缺陷检测+轮胎特征分析

1、展示轮胎图像，介绍其特征（**图和第二部分——预处理，分块处理共用一张图**）

2、讲一下为什么会属于分割任务（因为不同的缺陷面积会有不同的分级和处理方法）；引到自然图像的分割上。

（2）随着深度学习的流行，分割技术越来越多，大概 举例。

1、归归类，说说它们的做法有哪些不同（角度可以从感知多尺度方法入手。可以最后写）。

（3）这些方法针对自然图像有着很好的效果，但是由于轮胎数据的特殊性以及关注程度，出现的文章并不多，其中有\*\*方法。这些方法存在\*\*问题，我提出了一种\*\*，解决了\*\*\*。

The proposed method v2

前边introduction里应该说明原因，浅层学习到的东西是为后面的语义分割服务的，换句话说，浅层只是单纯的看得少，并不是为了小缺陷而存在的，因此就算融合了不同层结果依然有待提高。另一个池化层的存在也是为了语义分割服务的，同时不明显的小缺陷只会越来越小。再有，小缺陷检测不出来归根结底是因为小，我们把它们放大了看。

基于此，我们做了两方面的工作，一个是为检测小缺陷单独构建了网络，并使用

1. 总体框架图

结构，训练单独训练，之后再固定网络参数训练最后的融合层。测试时则是直接输出结果。

1. 粗分割
2. 细分割
3. 从问题入手：分析造成不精细的原因是什么（下采样）。下采样的作用：减少参数，提取更高的语义信息（前面介绍）。想提取细节信息一个最直接的方法就是不要下采样，但是不下采样参数规模会变大，很难训练。（借鉴于FSRCNN，一种对细节信息要求非常高的任务）。建立一个不含下采样的网络。
4. 没有下采样层的网络：
5. 在训练阶段，分块数据：为了进一步的缩小参数量，我们使用分块图像训练FSRCNN，一方面加速训练，另一方面可以有效地增加缺陷占比，维持缺陷特征与正常纹理之间的平衡，但是又没改变特征的纹理和周边像素的相对关系。测试阶段

The proposed method

（1）大体描述框架。（**图2，网络框架图**）

（2）预处理，分块处理（解释为什么轮胎图像可以分块，分块对比图）

（3）细的分割

1、解释一下为什么要用3层卷积，或者不用3层卷积用什么（可以最后写）。

2、分块之后怎么得到的

（4）粗的分割

借鉴了FCN的想法（暂时的，如果后面可以改成别的网络，但是重点我们只是借助于FCN的稠密预测能力，因为有细节提取网络，所以精细问题可以不用考虑）

实现细节

MULTI-SCALE DEFECT DETECTION NETWORK

这个工作旨在提高轮胎缺陷检测的精确度，特别是对边缘细节和小缺陷的检测。为了完成这个目标，一方面，我们的网络融合高级特征与低级线索通过两个并行的子网络。所提出的体系结构概述图如图2所示。语义感知网络具有挖掘高度抽象的语义特征包括形状和定位的能力，它简化自FCN，一种已经在自然图像的语义分割任务中有出色的表现的深度网络。进一步的，为了补充深度网络中过滤掉的细节信息，纹理感知网络被建立试图在不增加计算复杂度的情况下提取更多的浅层细节线索。另一方面，根据轮胎图像和子网络的特点，不同的数据与处理方式被使用去增强网络性能。在下一小节，我们详细描述了这些网络。

Our deep network aims to improve the accuracy of tire defect detection, especially for edge detail and small defects. To achieve this goal, on the one hand, the proposed MDDN combines deep layer features with low-level clues through two parallel sub-networks. Among them, the semantic-aware network is adopted to mine abstract semantic features including shape and position information, which is simplified by FCN[], a deep network that has outstanding performance in solving the semantic segmentation problems with natural images. Furthermore, in order to supplement the detailed information filtered during the extraction of high-level semantics, a novel texture-aware network is developed to attempt to preserve more shallow detail information without increasing computational complexity. On the other hand, owing to the characteristics of the tire images, a variety of data preprocessing methods are used to enhance enhance network performance for the two sub-networks. In the following subsections, we describe these networks and strategies in detail.

Semantic-aware sub-network

语义感知网络使用类似于众所周知的FCN网络的架构设计来捕获缺陷的高级语义。以VGG16为主要架构的FCN已经被证明在轮胎检测中是可行的。首先，卷积层和最大池化层被交替使用以获得图像中最具有代表性的信息，之后，VGG网络中的全连接层被特殊的卷积层替换以保留特征的空间信息。最后，与输入图像相同尺寸的预测结果被导出通过反卷积这些特征图。我们简化FCN网络成为一个二分类的网络，缺陷部分被视为目标，其他正常纹理视为背景。另外，我们规范化了卷积层参数，把pad和stride 设置为1，卷积和尺寸设置为3×3。这样的设置帮助我们去掉了原始FCN中的裁剪层，减少噪声的引入并且提高了检测速度，尽管这样会导致网络不能处理小于92×92的图像。

虽然VGG网络中the learned filters are very good generic visual descriptors，但是它最开始是为了目标分类任务，五个最大池化层的使用使其在拥有挖掘全局语义信息能力的同时过滤掉了不可忽视的细节信息，甚至是小尺寸的目标。因此以VGG16为主要架构的FCN对边缘细节不敏感。

Our deep network captures the desired high-level semantics of defects using an architectural design similar to the well-known FCNs [17], where the FCN with VGG16 as the basic framework has been proven to be viable in tire defect detection tasks. At first, a stack of convolution layers and max pooling layers are used repeatedly to obtain the most representative information in tire images. Each fully connected layer in the VGG16 is then replaced by a special convolutional layer to retain sufficient spatial information. Finally, prediction results with the same size as input images are derived by upsampling these spatial feature maps. We simplify an off-the-shelf FCN-VGG16 into a binary-classification and pixel-wise detection model. Moreover, we normalize the parameters of convolutional layers, such as both padding and stride are set to 1 and the convolution kernel size is set to 3×3. These parameters help us remove the crop layer in the original FCN, which reduces noise and increases the detection efficiency.

Although the learned filter in VGG network are excellent generic visual descriptors, it is originally trained for the purpose of object classification tasks. The existence of the five max pooling layers allows it to filter out the essential details while having the capability to mine global semantic information, especially for small-sized defects. Therefore, the FCN-VGG16 is not sensitive to the detection of edge details.

Texture-aware sub-network

如上所述，为了弥补由池化层过滤掉的细节信息，我们构建了一个提取浅层细节线索的网络，叫做纹理感知网络，与语义感知网络并行。受。。。的启发，在超分辨等对精准度要求较高的任务中有出色的表现。纹理感知网络使用4层卷积网络，特征被提取通过第一层卷积，带有5×5的卷积核。尽管这些特征可以被后面的结构使用，但是直接映射这些特征会导致计算复杂度增加，因此第二个卷积层尺寸被设置为1×1以节省计算花费和消除冗余，被叫做shrinking层。后面的两个3×3的卷积被用来提高特征表示能力。Without池化层的存在，细节信息能够被完全保留。Meanwhile，shrinking层在channel维度中的下采样能力既能够减少冗余信息又能保留图像的空间分辨率。

为了使纹理感知网络对尺寸变化更加鲁邦，我们使用分块和缩放做为数据预处理的策略。不同尺寸的图像块patches被裁剪自每一张原始轮胎图像做为网络的输入数据。自然图像中，需要首先推测出建议框，再对其进行精细化分割【】。而轮胎图像则可以直接规则裁剪，由于其背景纹理的相似性。裁剪后的图像中缺陷依然具有显著性，如图1b,c。本质上，裁剪和缩放操作改变了缺陷在正常图像的占比，使小缺陷变成了大缺陷。本文中，我们使用4块和16块的比例裁剪，0.5、1、1.5的比例缩放。

As mentioned above, to complement the details filtered by pooling layers, a novel texture-aware network, in parallel with the semantic-aware network, is developed to obtain detailed clues in shallow layers. The texture-aware network employs 4 convolutional layers inspired by FSRCNN, a network that is expert in handling precise visual tasks such as super-resolution. The first layer of convolution uses a 5×5 convolution kernel to extract the characteristics of the input image. These features are directly used by later layers to increase computational complexity. Although these features contain a wealth of detailed information, direct backward mapping increases the computational complexity. Therefore, the second convolutional layer uses a 1x1 convolution kernel to reduce computational cost and eliminate redundancy, called shrinking. The latter two 3x3 convolutions are adopted to improve representation capability. Without the existence of pooling layers, detailed features can be completely retained. Meanwhile, the down-sampling capability of the shrinking layer in the channel dimension can both reduce redundant information and retain the spatial resolution.

In order to make the texture-aware network robust to scale variations, images are preprocessed through multi-scale blocking and scaling strategies. Generally, in natural image segmentation tasks, the crop operation first obtains proposed regions, and then finely segmentation in these regions[]. In contrast, tire images can be cropped evenly while retaining defect semantic information, due to the similarity of background textures. As shown in Figure 1c, the cropping and scaling essentially change the proportion of defects in training images. In this paper, we crop 2×2 and 4×4 patches without overlap from each image, and consider scales of 0.5 and 1.5.

融合和训练

尽管纹理感知网络提取的特征中包含必要的细节信息，在没有全局语义的干预下，纹理感知网络的检测结果通常包含大量的噪声和误检。因此，融合语义感知网络和纹理感知网络是必不可少的。语义和纹理特征图直接被拼接和融合在channel维度上通过\*\*层和一个1×1的卷积层。

训练：由于两个子网络使用不同的数据与处理方式，因此我们首先使用分块数据训练纹理感知网络，然后训练整个网络带着固定的训练好的纹理感知网络参数。正如\*\*\*所表述的，由于轮胎图像纹理的特殊性，在测试阶段，原始图像可以被直接输入，并得到end-to-end的检测结果，实验证明了这种方式的有效性。

Although features extracted by the texture-aware network contain the essential details, the detection results derived from these features usually have a large amount of noise and a high false positive rate without the guidance of global semantics. Therefore, the fusion of semantic-aware networks and texture-aware networks is necessary. Semantic and texture feature maps are combined and fused in the channel dimension via a \*\* layer and a 1×1 convolution layer.

Since the two sub-networks use different input data and preprocessing methods, we first train the texture-aware network using the block data, and then train the entire network with fixed training texture-aware network parameters. As described in \*\*\*, owing to the characteristics of tire image textures, original images can be directly input during the test phase. And the end-to-end test results can be obtained. The experiment proves the effectiveness of this means.

Experiments

数据集和评价指标

消融实验结果比较

与现有方法比较

实验结果比较

和存在的轮胎图像的方法作对比

和其他的分割领域（用轮胎数据集fine-turn过）的方法对比

指标以及指标结果的表格

数据集

我们的实验数据集包含914张原始轮胎图像，其中既有胎冠图像又有胎侧图像，如图\*\*\*。这些图像包含各种各样的缺陷如。。。。。。对于语义感知网络，我们考虑翻转和镜像来增强数据。对于纹理感知网络，裁剪和缩放被采用，正如前面提到的。

Our experimental dataset consists of 914 tire images including both sidewall and tread images. These images involve various defects such as metal impurities, bubbles, and overlaps. For semantic-aware networks, we consider flipping and mirroring to enhance the data. For texture-aware networks, cropping and scaling are used, as mentioned above.

参数设置  
 。。。Cpu。。。Python。。。Caffe。 网络参数

Our proposed MDDN was coded with Python 3.5 in the Caffe framework. A GTX-1080 GPU and Intel Xeon-E5 3.40GHz CPU are used for both training and testing. The momentum parameter and weight decay were set to 0.99 and 0.0005 during training, respectively. Moreover, semantic-aware network is implemented on the public FCN code, and the parameters of the texture-aware network are randomly initialized.

评价指标。

We adopt the widely used metrics in instance segmentation community, including intersection over union(IOU) and pixel accuracy (PA). The former can represent the accuracy of position and area. And the latter indicates the ratio of the correct labeled pixels to the total pixels, which reflects the pixel-level accuracy of the detection results.

消融实验

1. 两个网络独立训练的结果和整体训练的结果比较
2. 测试时，用分块图像测试和原始图像直接测试的结果比较

In order to evaluate the parameter effectiveness of our proposed MDDN method, we conduct two groups of ablation experiments on the same data set. In one group, features learned by the semantic-aware network and the texture-aware network are used alone to detect defects. As the results shown in the figure \*\*. The semantic-aware network can roughly detect the location and shape of the defect. The texture-aware network has a high false positive rate due to the large amount of noise in the extracted features. In another group, for the same network trained with block images, we use the block images and the original images for testing. The experimental results verify that tire blocks still has relatively complete semantic information.

与现有方法的比较

FCN、FCN-16s、FCN-8s、SSD。。。

Conclusion

我们提出了\*\*\*，有效的解决了\*\*\*，实验结果证明我们方法的有效性。同时，哪些地方依然存在这问题。

In this paper, 我们提出了一个MDDN模型，被用于高效且精细的轮胎缺陷检测。通过融合语义感知和一个新奇的纹理感知网络，MDDN能在挖掘蕴藏在网络深层的语义信息的同时保留细节纹理。我们发现，缺陷区域在图像中占比的增加对小缺陷检测是必不可少的。另外我们还发现，通过分块策略能够有效地增强数据集，提升网络对细节纹理特征的感知。在相同的数据集下的实验结果表明，我们的MDDN与可比较的现存论胎缺陷检测方法相比有显着改进，并证明我们的方法可以产生更准确的小目标预测结果。未来的工作包括减少结构感知网络的噪音以及提高检测速度。

In this paper, we proposed a MDDN model for detailed tire defect detection tasks. Through combining a semantic-aware network and a novel texture-aware network, MDDN can preserve the necessary detail features while mining the semantic information hidden in deep layers. We showed that an increase in the proportion of defective areas is critical for small defect detection. In addition, we have experimentally verified that the blocking strategy can effectively enhance the dataset and retain detailed information, in tire images. The experiments demonstrate that our MDDN has significantly improved over the existing tire defect detecting methods, and can produce more accurate small defect detection results. The future work includes reducing noises in the texture-aware network and increasing detection speed.

0.0

需要阐明的问题：

1. 轮胎图像分块的可行性。一般图像处理没有人用分块，大多数采用推荐框的形式，缺点是一旦推荐框检测不出来，后续的网络就没法检测。我们用分块不是为了精确地找到缺陷，而是为了不丢失每一处细节，尽管它有可能不是缺陷。
2. 分块的好处。最大的好处是感知细节；间接增加了数据集，增大了有缺陷图像的比例，解决了检测没有缺陷的问题。
3. 可以加一个简单的实验，看看没有缺陷的小图像块有多少是真的检测出是没有图像的，说明对增强鲁棒性有一定的作用。
4. 一个网络感知细节信息，一个网络感知全局信息，两只眼睛看世界，具体最后的结果是谁做主导谁做辅助，也是靠训练来的。

9-23：

第二段：

Automatic defect detection, used to improve quality and accelerate production, has become an indispensable part in industrial processes \cite{kumar2008computer,li2016deformable}. Especially in tire manufacturing, numerous detection algorithms have been proposed \cite{zhang2013texture,zhang2018tire,xiang2014dictionary} and aroused extensive attention recently. In most real-world applications, tire defect detection is first carried out by deriving the defective region from tire X-ray images, which contains various types of defects caused by unclean raw materials and undesired manufacturing facilities \cite{guo2016defect}. Then, defective products are hierarchical processed according to the location and area of defects. Due to unique properties of the tire X-ray image, for instance complexity and low-quality, illustrated in previous studies \cite{zhang2013defect,wang2019tire}, most inspection processes are performed by human observers, which increases the risk and reduces the efficiency. Therefore, tire defect detection remains one of the most challenging inspection tasks.

轮胎缺陷检测重要，尽管已经有很多人研究它，但是依然是有。。我们针对。。

At present, existing detection methods are mostly devoted to distinguishing difference between defective regions and backgrounds (defective-free regions). A key issue for such methods is feature extraction. Early studies directly decomposed the defects using component analysis methods. Some existing hand-engineered filters are generally utilized to separate the texture and background components from inspected images, and defective regions are then derived from the residual image via adaptive thresholds. Obviously, the major shortcoming of these fundamental methods is the limitation of the information contained in low-level clues and spatial domain features. In order to take full advantage of the texture features as possible. Guo {\it et al}. \cite{guo2016defect} obtains an anomaly map of texture distortion by weighted averaging of the dissimilarity between each pixel and its neighbors, and detects the defects through segmenting this anomaly map. Besides, mulit-scale wavelet and curvelet transform are introduced in detection tasks, due to its capability of singularity analysis. Furthermore, optimized edge detection and total variation algorithm were used to achieve more accurate results \cite{yan2013detection}. Unfortunately, the representation capability of fixed kernels is not comprehensive enough. Moreover, the transform processes are computationally expensive. Recently, deep convolutional networks (ConvNets), which have outstanding performances in the recognition and segmentation tasks of natural images, were attempted to classify tire defects in \cite{cui2018tire}. With the excellent feature extraction capability of ConvNets, Wang {\it et al}. \cite{wang2019tire} further implemented the detection and segmentation in tire images by a fully convolutional network (FCN) \cite{long2015fully}. However, due to the existence of pooling layers, FCN is not sensitive to small defects and edge details, which is similar to that in dealing with natural image segmentation tasks.

早期的方法采用成分分解的方法[][][]，（介绍一下成分分解的步骤，两句）。Obviously, the major shortcoming of these fundamental methods is the limitation of the information contained in low-level clues and domain features. 为了充分地利用空间纹理特征，Guo提出了一种基于的。。。除此之外，多尺度小波和曲线波变换被介绍到轮胎缺陷检测任务在[][]，由于（小波的某种特性）。。。Furthermore, optimized edge detection and total variation algorithm were used to achieve more accurate results \cite{yan2013detection}. However, the representation capability of fixed kernels is not comprehensive enough. Moreover, the transform processes are computationally expensive.

第三段：

To achieve better detection performance on these small objects, many methods have been proposed in benchmark datasets.

Another fusing way can effectively address this problem by concatenating multi-scale features and detecting on top of the expanded feature maps, as shown in Fig. \ref{fig1}(b). For example, FCN defined a skip architecture to produce more accurate segmentation. Similar top-down skip architectures are popular in recent researches \cite{newell2016stacked,ghiasi2016laplacian}.

Since the training goal of the entire network is to extract abstract semantics, there exists a basic problem that it is still not enough to mine the detail texture in these structures \cite{zhou2018scale}. CrowdNet \cite{boominathan2016crowdnet} combined shallow and deep networks to overcome this shortcoming. Unlike in-network feature hierarchies, the shallow network is specifically designed to retain more details by reducing the number of pooling layers. Therefore, the capability to extract detail features is further enhanced.

The capability to extract detail features is further enhanced by a parallel shallow network with the deep network, rather than in-network feature hierarchies.

CrowdNet通过在浅层减少池化层来保留更多的细节信息，

第四段：

不同于CrowdNet，在纹理感知网络中，我们进一步将池化层减少到0，将细节和小缺陷信息完整的保留，最后的检测结果以语义感知网络为指导以过滤掉由于没有池化所带来的噪声。进一步地，为了获得更清晰地纹理细节，。。。

Different from the CrowdNet,

In texture-aware networks, pooling layers are discarded in order to completely retain detail textures and small-sized defects. Furthermore, image blocking......

9-24

MDDN

第一段

On the other hand, owing to the characteristics of tire images, a variety of data preprocessing methods are used to enhance performance for the two sub-networks. In the following subsections, we describe these networks and strategies in detail.

On the other hand, A variety of data preprocessing methods are used to enhance performance for the two sub-networks, owing to the characteristics of tire images, such as texture regularity in defective-free regions and diversity of various defects. In the following subsections, we describe these networks and strategies in details.

第二段

We simplify an off-the-shelf FCN-VGG16 into a binary-classification and pixel-wise detection model. Moreover, we normalize the parameters of convolution layers, such as both padding and stride are set to 1 and the size of convolution kernels are set to 3 $\times$ 3, and remove the crop layers to reduce noise and increase the detection efficiency.

Although the learned filter in VGG network are excellent generic visual descriptors, it is originally trained for the purpose of object classification tasks. The existence of the five max pooling layers allows it to filter out the essential details while mining global semantic information. Therefore, the FCN-VGG16 is not sensitive to the edge details detection, especially for small-sized defects.

We simplify an off-the-shelf FCN-VGG16 into a binary-classification and pixel-wise detection model. Specifically, both padding and stride are set to 1 and the size of convolution kernels are set to 3 $\times$ 3 in each convolution layer. The crop layers are removed to reduce noise and increase the detection efficiency. Although with these settings, small-sized input images cannot be processed after five pooling layers. In real-world applications, the size of the input x-ray image is fixed through the processing of x-ray devices and scaling operations.

第三段

纹理感知，表现在哪。

As mentioned above, to address this issue, a novel texture-aware network, in parallel with the semantic-aware network, is developed to obtain detailed clues in shallow layers. The texture-aware network employs four convolutional layers inspired by FSRCNN \cite{dong2016accelerating}, a network that is expert in handling precise visual tasks such as super-resolution. The first convolution layer uses 5 $\times$ 5 convolution kernels to extract the features of the input images. Although these feature maps carrying a large amount of valuable information can be directly mapped by next layers, this leads to an increase in computational complexity. Therefore, the second convolution layer uses a 1 $\times$ 1 convolution kernel to reduce computational cost and redundancy information, called shrinking layer. The latter two convolution layers with 3 $\times$ 3 kernels are adopted to improve representation capability. Without pooling layers, detailed features can be completely retained. Meanwhile, the down-sampling capability of the shrinking layer in the channel dimension can both reduce redundant information and retain the spatial resolution.

为什么会丢失细节，因为池化层（加一个折线图表示细节语义与池化层层数间的关系）

把池化层直接扔掉行不行，不行，为什么，

怎么解决这个问题，构造一个浅层的只有卷积层的网络。

分别介绍网络的每一层。

1. （先介绍为什么会丢失细节信息）如前所述，池化层的使用使得网络能够获取准确的语义信息，同时，一些有用的细节信息被不可避免的过滤掉。以Fcn-vgg16为例，图\*\*展示了池化层对特征提取的影响。但是从现有的深度网络中直接丢弃池化层会带来一系列的训练难题，例如参数爆炸，过拟合，【crowdnet】。因此我们特意设计它变浅且只包含4层卷积层，并且设置shrinking层间接实现降维。类似的结构被使用在[][]中，解决超分辨问题。一方面，

语义感知网络不需要通过很深的深度获取精确地语义信息。另一方面，the down-sampling capability of the shrinking layer in the channel dimension can both reduce redundant information and retain the spatial resolution.

Since blob detection does not require the capture of high-level semantics, we design this network to be shallow with a depth of only 3 convolutional layers.

As mentioned above, with several pooling layers, especially the max-pooling layers, deep networks can extract high-level abstract features and semantic information, while detail clues are unavoidable filtered. Taking Fcn-vgg16 as an example, Figure \*\* indicates the impact of pooling layers on the extracted features. However, directly dropping pooling layers from existing deep networks will bring a series of training problems, such as parameter explosion and overfitting. Therefore, our shallow texture-aware network was cautiously designed with only four layers of convolutional layers, where the shrinking layer is utilized to indirectly reduce the parameter dimension. A similar structure is used in FSRCNN \cite{dong2016accelerating} to handling precise visual tasks such as super-resolution. On the one hand, the representation of the texture does not require the capture of high-level semantics. On the other hand, the down-sampling capability of the shrinking layer in the channel dimension can both reduce redundant information and retain the spatial resolution. As shown in \*\*, the first convolution layer uses 5 $\times$ 5 convolution kernels to extract the features of the input images. Although these feature maps carrying a large amount of valuable information can be directly mapped by next layers, this leads to an increase in computational complexity. Therefore, the second convolution layer uses a 1 $\times$ 1 convolution kernel to reduce computational cost and redundancy information, called shrinking layer. The latter two convolution layers with 3 $\times$ 3 kernels are adopted to improve representation capability. Without pooling layers, detailed features are completely retained through the texture-aware network.

第四段

Combination

Although features extracted by the texture-aware network contain the essential details, the detection results derived from these features usually have a large amount of noise and a high false positive rate without the guidance of global semantics. Therefore, the fusion of semantic-aware networks and texture-aware networks is necessary. Semantic and texture feature maps are concatenated and fused in the channel dimension via a {\it concat} layer and a convolution layer with 1 $\times$ 1 kernels. Considering the semantic network as a guide and texture network as a supplement, final predictions are automatically learned by training an additional convolution layer.

第五段

Pre-processing and training

In order to make the texture-aware network more robust to scale variations, input images are preprocessed through multi-scale blocking and scaling strategies. Generally, in natural image segmentation tasks, the blocking operation first obtains object proposals \cite{girshick2014rich,ren2015faster}, and then finely segment in these regions. In contrast, tire images can be cropped evenly while retaining defect semantic information, due to the similarity of background textures. As shown in Figure 1c, the cropping and scaling essentially change the proportion of defects in training images. In this paper, we crop 4 and 16 blocks without overlap from each image, and consider scales of 0.5 and 1.5.

Since the two sub-networks use different input data and preprocessing methods, we first train the texture-aware network using the block data, and then train the entire network with fixed texture-aware network parameters. As described in the previous subsections, owing to the characteristics of tire image textures, original images can be directly fed to obtain end-to-end results during test phases. The feasibility of this strategy are proved via next ablation experiments.