Main-Secondary Network for Defect Segmentation of Textured Surface Images

Yu Xie
Fudan University
xieyu9332@gmail.com

Fangrui Zhu
Fudan University
xiaoruirui233@gmail.com

Yanwei Fu[1]
Fudan University
yanweifu@fudan.edu.cn

Abstract

Building an intelligent defect segmentation system for textured images has attracted much increasing attention in both research and industrial communities, due to its significance values in the practical applications of industrial inspection and quality control. Previous models learned the classical classifiers for segmentation by designing hand-crafted features. However, defect segmentation of textured surface images poses challenges such as ambiguous shapes and sizes of defects along with varying textures and patterns in the images. Thus, hand-crafted features based segmentation methods can only be applied to particular types of textured images. To this end, it is desirable to learn a general deep learning based representation for the automatic segmentation of defects. Furthermore, it is relatively less study in efficiently extracting the deep features in the frequency domain, which, nevertheless, should be very important to understand the patterns of textured images. In this paper, we propose a novel defect segmentation deep network – Main-Secondary Network (MS-Net). Our MS-Net is trained to model both features from the spatial domain and the frequency domain, where wavelet transform is utilized to extract discriminative information from the frequency domain. Extensive experiments show the effectiveness of our MS-Net.

1. Introduction

With the advancement of automation in manufacturing, automation of material quality inspection with little human intervention is in high demand. To meet the industry standards and guarantee stringent quality control limits, it is of great significance to carry out product inspection in advance and put rejected products in wasted upstream factory capacity. In industrial production line, there are various kinds of surface defects, such as metallic defects [32], railway track defects [20] and mobile screen defects. Defects can be tiny, fine-grained, which are unrecognizable by humans and extremely hard for intelligent inspection systems to detect. Generally, the defect inspection can be framed as three tasks, *i.e.*, defect images classification, defect part detection and defect part segmentation.

Several excellent previous works [30, 33, 20, 21, 41] have studied the task of image based quality inspection, and these works generally focus on either classification or detection. The defect segmentation is much more challenging than defect classification and defect detection, since it needs to determine the defective region pixel-by-pixel. Moreover, accurate and precise defect segmentation results are critical for evaluating the product quality. For example, one important measure for manufacturer in judging whether one mobile screen is of good quality, is the size of total defective regions. It is thus essential to automatically and efficiently learn to segment the defects. Despite remarkable efforts have been made in industrial defect segmentation, it still remains as a challenging problem.

Lack of Adaptability. In the past years, there are some machine vision based methods [4, 22] to perform defect segmentation, *e.g.*, SVM based method [29], random forest based method [26] and so on. These methods usually design different hand-crafted features for particular tasks, and classifiers are then trained on these features. In that case, elaborately designed features can fit well under particular conditions, but may fail to effectively segment defects of other industrial surfaces. Thus, these methods are lack of adaptability and cannot be directly adapted to various conditions, which is often the case in product quality control system. To this end, training a defect segmentation model automatically and generalizing the model to different cases are in high demand.

Ambiguous Defective Regions as well as Changing Textures. In some cases, defects look very subtle and unrecognizable due to uneven illumination as shown in the two row of Fig. 1. Although these defects are slight, they have a great impact on the use of the product. In industrial product inspection, even tiny and sparse defects shown on the products can potentially lead to a catastrophic failure of the whole system, and such products should be abandoned.

[1] Y. Fu is the corresponding author of this article . He is with School of Data Science, and MOE Frontiers Center for Brain Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University

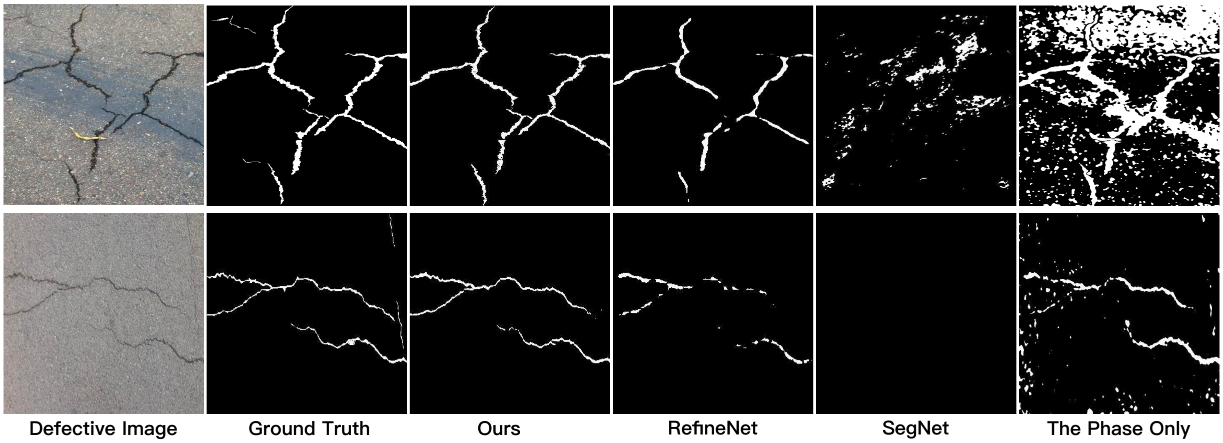

Figure 1. Defect segmentation by different methods. The first column shows the original defective images containing recognizable and unrecognizable defects. The second column shows the ground truth where defects are annotated by experts with particular criteria. From the third to the last column, segmentation results by different methods are compared, which demonstrate that our model can segment unrecognizable defects and produce fine-grained results better than other methods.

Generally speaking, it is often the case that there are very similar visual patterns between the defective regions and background textures. Background textures are often changing in either color distribution or illumination, which may appear different from the defect patterns. Such the visual similarity may confuse most of the state-of-the-art defect segmentation methods, which are often based on features of spatial domain to detect defects. Therefore, it is essential to construct a defect-sensitive model, which can recognize and separate not only obvious but also ambiguous defects from various textures.

High Accuracy Requirements. For the purpose of industrial surface inspection, it is necessary to know the exact size and pixel-wise location of the defect, as illustrated in Fig. 1. Such results are regarded as an important indicator to evaluate the quality of a product. Most previous works can only detect defects with bounding boxes rather than pixel-wise defect measures, bringing difficulties in performing remedies in real applications as well as the need in further inspection tools. Therefore, pixel-level segmentation for defects is essential in maintaining safety and quality control. In that way, the accuracy of defect detection can also be improved. However, the performance of existing segmentation works is still quite limited in achieving the pixel-level accuracy.

Currently, deep learning [14], especially the convolutional neural network (CNN), has achieved remarkable performance in many computer vision applications, as it enables the model to be learned automatically and takes advantage of large amounts of data to gain high precision. In this paper, we explore a novel deep learning approach for the defect segmentation. Different from the traditional CNN only extracting features from the spatial domain, we propose a novel *Main-Secondary Network* (MS-Net) to segment defects by using features extracted from both spatial domain and frequency domain. Particularly, we use the convolutional layers to extract the features from spatial domain and Discrete Wavelet Transform (DWT) to decompose the image to extract features from frequency domain. Thus, the whole network is composed of two sub-nets, *i.e.*, Main Net and Secondary Net. These two sub-nets jointly learn defect features from two domains and features are integrated for the segmentation.

This is inspired by different components containing in the frequency domain. Since defective regions often appear with some disparities and unsmoothness in the image, combining components from two domains can leverage patterns of defects and textures under different conditions. Wavelet transform (WT) has been shown to be an efficient tool to depict the contextual and textural information of an image at different levels [12], motivating us to incorporate WT to a CNN-based defect segmentation system. In particular, we use convolutional layers to construct a Main Net not only for extracting spatial features but also learning visual embeddings of wavelet coefficients. The Secondary Net is adopted to obtain features from frequency domain by taking advantage of wavelet transform, which is often considered as a powerful tool for approximating arbitrary nonlinear functions, appropriate for ambiguous defects.

As shown in Fig. 1, our model produces fine-grained segmentation results with high precision and beats other state-of-the-art segmentation methods.

Contribution. Main contributions are listed as below: (1) We propose a novel network structure for surface defect segmentation that can automatically detect and segment ambiguous defects from various background textures and produce fine-grained results. (2) We provide a new strategy in combining frequency and spatial domain features for computer vision and image processing tasks by utilizing discrete wavelet transform and its inverse process. Experimental results show that our proposed approach performs well.

2. Related work

Industrial Surface Inspection. Generally, industrial surface inspection tasks are concerned with identifying defective regions that deviate from background textures based on certain criteria. To design particular criteria for different problems, several methods have been proposed by using hand-crafted features.

Statistical measures [9, 39] are used for identifying defects from textures, such as histogram statistics, local binary patterns and so on. Structural approaches [36] characterize textures by *texture primitives* and detect defects by generalizing the spatial placement rules based on those primitives. There are also model based methods such as random field model [15], the *texem* model [40]. Filter based approaches [33, 1] often apply filter banks on the image and compute the energy of the filter response. In general, these works employ the hand-crafted features to train classical classifiers such as SVM for predicting pixel-wise defects from industrial images.

However, they are prone to errors due to the complexity and variation of background textures and they often have high cost in time. They may perform well under particular conditions, but fail in other scenarios, inappropriate in real industrial applications for the lack of adaptability. Current deep learning based methods [21, 13] have also been proposed, but they are mainly for defect detection. Different from above methods, we present a deep segmentation framework, which can be utilized under different conditions and generalized as a whole. Our method wraps different kinds of features and trains the network in an end-to-end manner.

Defect Segmentation. With the development of convolutional neural network (CNN) for image based tasks, several defect segmentation networks [34, 32, 27] have been designed to tackle the adaptability problem of hand-crafted features. For example, Xian *et al.* presented an inspection system for metallic surface, consisting of defects detection and classification in coarse-to-fine manner. Since most of existing works are for defect detection [41, 21] rather than pixel-wise segmentation, defect segmentation tasks remain very challenging. Although defects can be unrecognizable and ambiguous along with variations in background textures, we propose a method performing well on diverse datasets.

Object Instance Segmentation. Driven by the recent success of R-CNN [8], the task of object instance segmentation has been widely developed [24, 35], which is defined as per-pixel object classification of the image. It is often referred as semantic segmentation, which is quite different with defect segmentation task in that: (1) Image pixels are classified into a series of categories for semantic segmentation, while defect segmentation is considered as a binary classification problem on image pixels. (2) There are often many semantic information and object features extracted by elaborate deep networks, such as U-Net [25], Mask R-CNN [10], which are relatively less in defective images. (3) The foreground defective regions are often overlapped with the background textures without obvious semantic deviation, making the defect segmentation extremely difficult. Furthermore, due to different objectives of the two tasks, state-of-the-art semantic segmentation methods usually don't work in defect segmentation tasks.

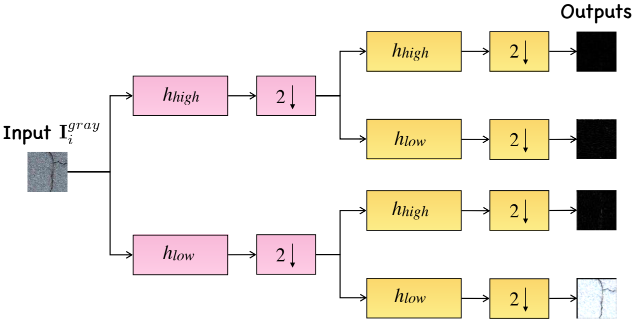

Figure 2. Illustration of Discrete Wavelet Transform (DWT). The 2D Haar wavelet transform on the grey-scale defective image $\mathbf{I}_i^{gray}$ can be shown from the input to outputs. Firstly, one-dimensional DWT is performed on the rows to get the low-pass and high-pass frequency components of the input image. Secondly, another one-dimensional DWT is used on the columns to compute final wavelet coefficients. The inverse wavelet transform (INVWT) is just the opposite from DWT. It is to reconstruct the grey-scale image from four wavelet coefficients.

Wavelet Transform. Wavelet Transform (WT) has been an efficient tool in image super resolution [12], image restoration [17] and also defect detection / segmentation [37, 2, 38]. Wong et al. [38] propose a a stitching detection and classification technique, combining the improved thresholding method based on the wavelet transform with the back propagation (BP) neural network. Wen et al. [37] use wavelet transform (WT) and a co-occurrence matrix (CM) to extract features of texture images, then use those features to locate defects on textile fabrics. Arivazhagan et al. [2] propose to utilize Gabor wavelet transform to detect the defects in fabrics. Different from them, we use WT to transform the image to extract its information in frequency domain and introduce WT into CNN layers, combining both spatial and frequency domain features for segmentation.

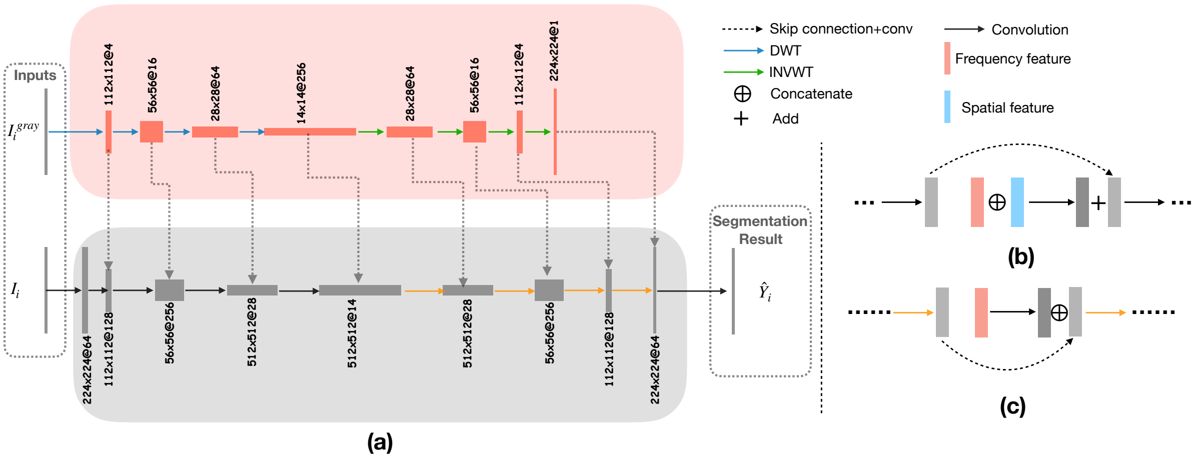

Figure 3. MS-Net Architecture. This figure shows the detailed structure of our MS-Net, where the Secondary Net embeds DWT into CNN to obtain features of frequency domain and skip connections are used to integrate features from two sub-nets. (a) The overall structure of the network.(b) shows how the network combines the frequency domain feature and the spatial domain feature at the encoder stage. (c) shows how the network combines two features at the decoder stage. The inputs of two sub-nets are the original defective image $\mathbf{I}_i$ and its corresponding grey-scale image $\mathbf{I}_i^{gray}$, and our model outputs the segmentation result, which is a grey-scale image $\hat{Y}_i$. The height, width and channel-size of the feature map are shown for each layer.

3. Methodology

3.1. Model Overview

We propose a novel deep network called Main-Secondary Network (MS-Net). As shown in Fig. 3, our MS-Net is composed of two parts, *i.e.*, Main Net and Secondary Net. Particularly, the Secondary Net is used to extract and process the features of defective images in the frequency domain. The Main Net plays two roles. On the one hand, it extracts features of input images in the spatial domain by convolutional layers. On the other hand, it consequently fuses the multi-channel features from the Secondary Net and combines them to feed into convolutional layers. The final segmentation result is obtained from the output of the Main Net.

3.2. Problem Definition

Given the defective image set D, we aim to train a pixel-wise classifier to predict the defects. Particularly, we have $D = \{\mathbf{I}_i, Y_i\}_{i=1}^{N}$, where $Y_i = \left\{y_m^{(i)}, m = 1, ..., |\mathbf{I}_i|, y_m^{(i)} \in \{0,1\}\right\}$ denotes the ground truth annotation / segmented result. For each image, it has a pixel-wise label with $y_m = 1$ representing a defective pixel and $y_m = 0$ representing a non-defective pixel. The goal is to capture the conditional distribution of defect pixels and learn the mapping function:

$$\hat{Y}_i = M_{net}\left(\mathbf{I}_i, S_{net}\left(\mathbf{I}_i^{gray}\right)\right) \tag{1}$$

where M_{net} denotes the Main Net and S_{net} denotes the Secondary Net. $\hat{Y}_i$ is the segmentation result.

3.3. Defect Segmentation

Secondary Net. As shown in Fig. 3, the Secondary Net is an autoencoder structure with a down-sampling path and an up-sampling path. In order to extract features from the frequency domain and combine them with CNN, we design a new operation block by using Discrete Wavelet Transform together with convolutional layers (DWT + Convolution).The frequency domain features are passed to the Main-Net after passing through the convolution layer. The Discrete Wavelet Transform (DWT) is often used to capture both frequency and location information of feature maps [7, 17], which may be helpful in preserving texture details. Since DWT is invertible, all information can be kept through DWT and its inverse process (INVWT).

The Secondary Net takes a grey-scale image $\mathbf{I}_i^{gray}$ of the size $h \times w \times 1$ as input and represents it as a set of feature maps. Particularly, this block first use the Haar based DWT to transform and map the input to its frequency distribution. We choose Haar based DWT for it is easy enough to operate and depict defect information of different frequencies and it is appropriate in terms of speed. The details are shown in Fig. 2, where the input grey-scale image is decomposed into four quarter-sized images, the inner products of the original image and different wavelet basis functions with interval sampling in the end.

After each DWT, the transformed image passes through ResNet [11] basicblock based convolutional layer to learn a compact representation as the inputs to the subsequent DWT. The number of feature maps (or the channel-size) increases in the down-sampling path to explore enough information from DWT. During the down-sampling process, features with the same channel size are concatenated with the Main Net feature maps and the image is encoded as different embeddings. Skip connections are used to concatenate feature maps.

In the up-sampling stage, the image is reconstructed by INVWT and passes through ResNet [11] basicblock based convolutional layers to extract additional features from the result of INVWT. As shown in Fig. 2, the inverse process is from the righthand side to the lefthand side, where the channel-size will reduce to one quarter of the former feature map. Similarly, features extracted from the INVWT will concatenate with Main Net feature maps, by which the frequency characteristics of DWT are expected to benefit the defect segmentation. As mentioned above, the overall Secondary Net has 8 blocks.

The overall process can be summarized as :

$$S_{net}\left(\mathbf{I}_i^{gray}\right) = f_{IC}\left(f_{DC}\left(\mathbf{I}_i^{gray}\right)\right) \tag{2}$$

where f_{IC} indicates the INVWT operation in the up-sampling process and f_{DC} denotes the DWT operation in the down-sampling process.

Main Net. As shown in Fig. 3, the Main Net is also structured as an autoencoder, where we take the original image $\mathbf{I}_i$ with size $h \times w \times 3$ as the input and the output is the segmentation result with size $h \times w \times 1$. The encoder down-samples the input to get increasing number of feature maps from the spatial domain, as well as combining frequency domain features consequently. The decoder reconstructs the image to the same spatial size with the input image but containing only one channel. All the convolutional filters are in kernel size 3×3 or $1 \times 1$with a stride of 2 or 1 and each convolution is followed by a ReLU. We use the skip connection to concatenate features between the Main Net and the Secondary Net, so that features from both domains are utilized for defect segmentation. There are some convolutional layers between fusion feature layers to further obtain visual embeddings of defects and textures. Besides, to prevent gradients vanishing and accelerate convergence, we mainly refer to the convolutional layer structure of resnet.

Loss Function. This network is trained in an end-to-end manner. The backward process is simple yet effective. We use the cross entropy loss on the output of the Main Net and back-propagate the loss through the two sub-nets to update model parameters. The loss function can be written as:

$$\mathcal{L}_{seg} = Y_i log\left(\hat{Y}_i\right) + (1 - Y_i)\, log\left(1 - \hat{Y}_i\right) \tag{3}$$

4. Experiment

4.1. Datasets and Settings

We evaluate the proposed method on three datasets, *i.e.*, DeepCrack [18], CrackForest [6, 28]. We highlight that our framework is able to segment defects of various complex surface images.

CrackForest Dataset [6, 28]. It contains annotated road crack images which can reflect urban road surface condition in general. The total number of images is 118, with hand labeled ground truth contour for each image. The road defect dataset is more complex because it is close to natural scenes. In addition to defects such as cracks, there are also gasoline spots and so on, posing challenges to the robustness of the method.

DeepCrack Dataset [18]. It is a benchmark dataset with cracks in multiple scales and scenes to evaluate the crack detection systems.DeepCrack is a road crack data. The difference from CrackForest data is that deepcack has a larger amount of data, more noise in the data, and more difficult segmentation.All of the crack images in this dataset are manually annotated. It Contains a total of 300 training data, including pictures of different sizes and a total of 237 test images with different sizes.

4.2. Evaluation Protocols

We adopt four commonly-used metrics [19] for semantic segmentation and scene parsing evaluations. Let n_{ij} be the number of pixels of class i predicted to belong to class j, where there are n_{class} different classes, and let $t_i = \sum_j n_{ij}$ be the total number of pixels of class i. Here, we have only two classes (defective and non-defective) and thus $i = 2$. The four metrics are:

(1) Pixel Accuracy (pAcc): $\sum_i n_{ii} / \sum_i t_i$, the percent of pixels in the image which are correctly classified.

(2) Mean Accuracy (mAcc): $(1/n_{cl}) \sum_i n_{ii}/t_i$, the average on pixel accuracy values of all classes.

(3) Mean IoU (region intersection over union) (mIoU): $(1/n_{cl}) \sum_i n_{ii}/(t_i + \sum_j n_{ji} - n_{ii})$, the average on IoU (the percent of overlap between the target and the predicted output) of all classes.

(4) Frequency Weighted IoU (f.w.IoU): $(\sum_k t_k)^{-1} \sum_i t_i n_{ii}/(t_i + \sum_j n_{ji} - n_{ii})$, mean IoU with weights of each class.

4.3. Implementation Details

We use PyTorch [23] for all experiments. Our model is trained with one TITAN Xp GPU and takes about one hour to complete training. In the training stage, defective images and their corresponding segmentation annotations are used to train the network from scratch. The inputs of Main Net and Secondary Net are original color images and their corresponding grey-scale images, respectively. We use the SGD optimizer to train the model and our model gets converged after 150 epochs on all datasets. The momentum is set to 0.9. The initial learning rate is set to 0.1 and decays by 10^{-1} after 15 epochs.

In the data pre-processing stage, we perform data augmentation on the DeepCrack dataset and the CrackForest dataset by following the commonly-used method [5, 35] in medical image processing. Since these two datasets only contain a few defective images, in order to train the model better and make it robust, we randomly crop large-size images into multiple 224×224 small-size images while maintaining the structure of them. Thus, our model can be fully trained and maintain the robustness as well.

4.4. Competitors

We compare several competitors here that potentially can be used to segment defects, including the learning based approaches and hand-crafted features based methods. (1) PHOT [1]: This method is a filter based method. It uses Fourier transform and Gaussian filtering to extract the frequency domain knowledge of the image, which differs from our method in that it only uses frequency domain features and is an un-learnable method. (2) SegNet [3] and U-Net [25] : SegNet and U-Net is a semantic segmentation network with auto encoder structure, Feature extraction through encoder and semantic segmentation using decoder. (3) RefineNet [16]: RefineNet is a generic multi-path refinement network that explicitly exploits all the information available along the down-sampling process to enable high-resolution prediction using long-range residual connections. RefineNet, U-Net and SegNet both use deep networks to achieve segmentation. But unlike our approach, they only use spatial domain features and does not use frequency domain features.

Note that recent state-of-the-art segmentation methods including Mask R-CNN [10] have also been investigated to be utilized to segment the defects. However, we find that they are quite difficult to either train from scratch or be fine-tuned over above datasets, mainly due to the model complexity and the intrinsic differences between objective images. In the experiments, recent segmentation methods have shown poor results and thus do not appear here as competitors.

4.5. Results Analysis

We analyze our experimental results from both quantitative and qualitative aspects.

Quantitative Results. We demonstrate the application of the proposed MS-Net on the test sets of two datasets. Table 1 lists the results of the competing methods on these two datasets, where our method has outperformed state-of-the-art methods by all metrics in general. Our method beats the method based on frequency domain features only and deep learning based method. By comparing to results of U-net [25], SegNet [3], RefineNet [16] and PHOT [1], we can demonstrate the effectiveness of our MS-Net in exploiting discriminative features from two domains (spatial domain and frequency domain) and applying in the defect segmentation task, since SegNet and RefineNet mainly take advantage of spatial domain features and PHOT utilizes frequency domain features.

We note that our MS-Net generally achieves favorable performance when compared with the competing methods. This suggests the overall performance advantages of the proposed MS-Net in the capability of multi-source (spatial domain and frequency domain) information extraction and fusion for adapting the defect segmentation model. When compared to the existing methods of defect segmentation mIoU, the performance margins are quite large. This indicates the importance of learning from both domains on labeled defect images, since hand-crafted features are not sufficiently generalizable across different datasets with varying texture and defect conditions. Note that those metrics in 4.2 are mostly used for semantic segmentation evaluation, among which mIoU improvement can directly suggest the model design advantages of our MS-Net in exploiting the diverse knowledge in ambiguous defects and changing textures, typical in real industrial deployments.

It is worth noting that the performance advantages by our MS-Net are achieved using only a few layers, which is quite applicable in real industrial inspection scenarios.

Qualitative Results. The visualization of the segmentation results of our MS-Net as well as the competing methods are shown in Fig. 4, 5.

As in Fig. 4, the segmentation results show that our MS-Net can output fine-grained segmentation results with sharp separation between defects and the texture. The defects on road images appear to be winding through the whole image and show much unsmoothness along the way, extremely hard to achieve accurate pixel-wise segmentation results. Our MS-Net can segment those kinds of defects quite well in terms of sharpness and granularity, while other methods can hardly produce sound and clear segmentation results.

For ambiguous defects of industrial images, our model can also segment them quite correctly as shown in Fig. 4. Visual comparisons of the competing methods on these two datasets indicate the robustness and promising performance of our MS-Net. The defects on CrackForest are quite very thin and can be successfully segmented by deep learning based methods, but our MS-Net performs better with respect to pixel accuracy and comprehensiveness. The appearance of the defect is closer to the texture and they are easily confused with those of image texture in general, leading to some weakness for other methods. This suggests that our MS-Net can successfully discriminate the pattern difference between texture and sharp defects by taking both spatial domain features and frequency domain features into consideration.

More complex texture changes are shown in Fig. 5, In the deep crack dataset, the texture is more complicated,More

	DeepCrack [18]				CrackFroest [28]			
	pAcc	mAcc	mIoU	f.w.IoU	pAcc	mAcc	mIoU	f.w.IoU
SegNet [3]	0.930	0.716	0.574	0.896	**0.984**	0.500	0.492	**0.970**
RefineNet [16]	0.979	0.840	0.783	0.961	0.980	0.657	0.570	0.967
Unet [25]	0.979	0.876	0.7962	0.962	0.981	0.692	0.596	**0.970**
PHOT [1]	0.658	0.810	0.392	0.622	0.895	0.021	0.014	0.867
Ours	**0.986**	**0.914**	**0.849**	**0.974**	0.980	**0.799**	**0.626**	0.968

Table 1. Quantitative Results of MS-Net on 'DeepCrack', 'CrackForest'. We compare our method with both traditional image processing method and deep learning based methods.

likely to cause textures to be identified as defects,Our network reduces the number of false detections by combining different features and segmented from the such images, showing the detection capabilities of our model in complex environments.

Thanks to the frequency characteristics of DWT and its combination with spatial domain features, our MS-Net can correctly detect diverse defect patterns on various textures and produce fine-grained and high-precision segmentation results. Furthermore, we highlight that our method can not only be applied on images for visual inspection and quality control, *e.g.* DeepCrack [18] and CrackForest [6, 28] datasets, but also has effects on other types of images, We did some experiments on other types of datasets in the ablation study.

4.6. Ablation Studies

In order to investigate the effectiveness of our model elements for the defect segmentation task, we conduct several ablation studies to measure the importance of two major elements, *i.e.*, spatial domain features and frequency domain features.

Only Frequency Domain Features (Secondary Net). The main benefit of our model is assumed to have combined features from frequency domain by utilizing DWT. Other than previous methods such as PHOT [1], we adopt the deep learning framework and embed wavelet transform to obtain frequency domain knowledge. In order to verify the significance of this operation, we conduct the defect segmentation by using only the Secondary Net of the MS-Net. Although the Secondary Net combines DWT with convolutional layers, it can be regarded as the main use of features from frequency domain and convolution operations are utilized to extract more knowledge from DWT. We have done experiments on the two datasets with the same training configurations as in Sec. 4.3. The quantitative results are shown in Tab. 2.

Only Spatial Domain Features (Main Net). Since most of previous methods are based on learning features of spatial domain and they can achieve high performance for semantic segmentation tasks and so on, we conduct experiments to test the results of only using Main Net for segmentation, where only spatial domain features are taken into considerations. In that case, features from the frequency domain by skip connections are omitted and the segmentation model only contains a few layers. The training configurations are

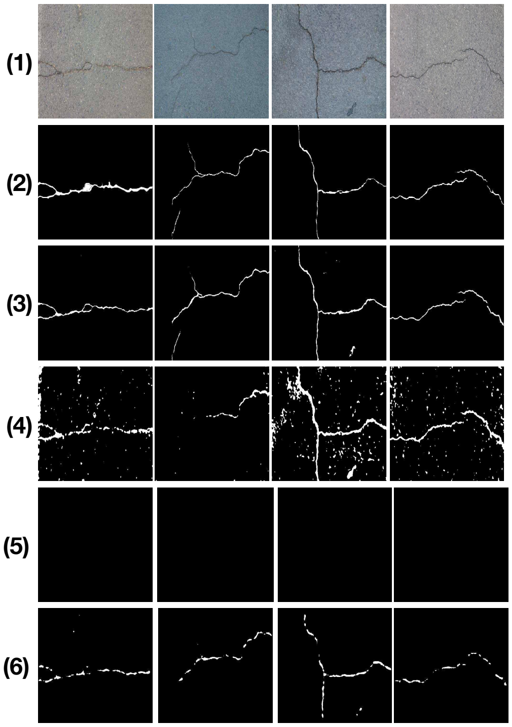

Figure 4. Qualitative Results on CrackForest by Different Methods. The meaning of each row: (1) The original images in the datasets. (2) Ground truth annotations of defects. (3) Segmentation results of our MS-Net. (4) Segmentation results of PHOT. (5) Segmentation results of Seg-Net. (6) Segmentation results of RefineNet.

	pAcc	mAcc	mIoU	f.w.IoU
CrackForest [6, 28]	0.839	0.606	0.514	0.791
DeepCrack [18]	0.956	0.500	0.478	0.916

Table 2. Quantitative Results of Ablation Study on Only Using the Secondary Net.

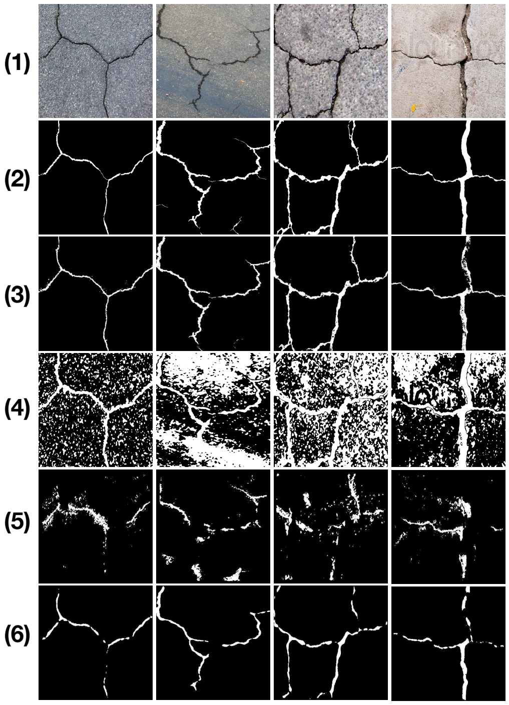

Figure 5. Qualitative Results on DeepCrack by Different Methods. The meaning of each row: (1) The original images in the datasets. (2) Ground truth annotations of defects. (3) Segmentation results of our MS-Net. (4) Segmentation results of PHOT. (5) Segmentation results of Seg-Net. (6) Segmentation results of RefineNet.

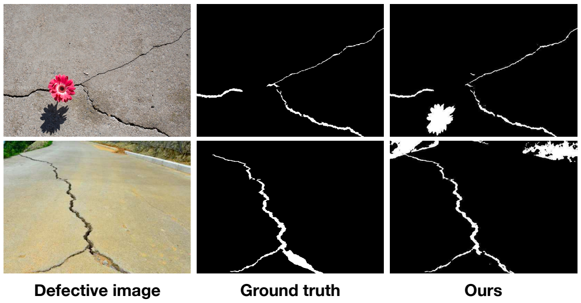

Figure 6. Some failure cases of MSNet. The meaning of each column: (1) The original images in the datasets. (2) Ground truth annotations of defects. (3) Segmentation results of our MS-Net.

the same with Sec. 4.3. The quantitative results on three datasets are shown in Tab. 3, which are not good as combining features from two domains.

As shown above, the segmentation performance can be improved after combining the two sub-nets. The likely reason for this is that combining DWT with CNN can boost the feature separation of defects from background textures by decomposing the image to several channels in the frequency domain.

We also show some failure cases. From the failure cases, we can see that the network identified the shadow of the object and the grass on the roadside as defects.The reason for this result is on the one hand because there is very little data in this case in the training dataset, and it may also be that the network tends to learn low-level features after combining the frequency domain features.

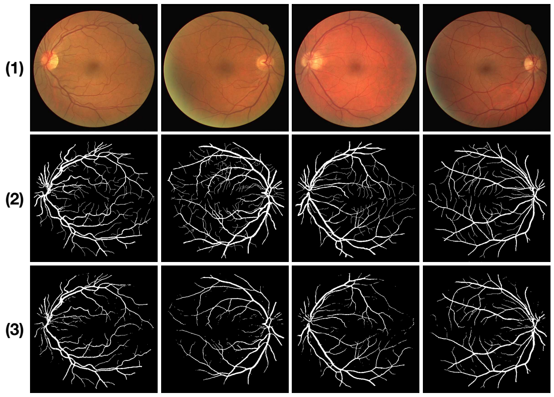

Figure 7. Some failure cases of MSNet. The meaning of each column: (1) The original images in the datasets. (2) Ground truth annotations of defects. (3) Segmentation results of our MS-Net.

	pAcc	mAcc	mIoU	f.w.IoU
CrackForest [6, 28]	0.977	0.659	0.558	0.966
DeepCrack [18]	0.973	0.844	0.744	0.954

Table 3. Quantitative Results of Ablation Study on Only Using the Main Net.

We have also performed experiments on other types of datasets to verify the validity of our model. We performed a blood vessel segmentation experiment on the drive dataset [31], Visualization results are shown in Fig. 7.It can also work well on the drive dataset [31].

5. Conclusion

In this work, we propose a novel MS-Net framework for defect segmentation, which will benefit industrial surface inspection and product quality control. To the best of our knowledge, we are the first to jointly utilize features from frequency domain extracted through DWT and features from spatial domain by CNN for separating defects from different background textures. We further analyze our model components. Most importantly, our method outperforms state-of-the-art methods in terms of accuracy and speed. In the future, we will investigate extending our framework to more segmentation cases and other related tasks.

6. Acknowledgments

This work was supported in part by Science and Technology Commission of Shanghai Municipality Project(#19511120700,#19ZR1471800).

References

[1] D. Aiger and H. Talbot. The phase only transform for unsupervised surface defect detection. In *Emerging Topics In Computer Vision And Its Applications*, pages 215–232. World Scientific, 2012.

[2] S. Arivazhagan, L. Ganesan, and S. Bama. Fault segmentation in fabric images using gabor wavelet transform. *Machine Vision and Applications*, 16(6):356, 2006.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[4] F. G. Bulnes, R. Usamentiaga, D. F. Garcia, and J. Molleda. An efficient method for defect detection during the manufacturing of web materials. *Journal of Intelligent Manufacturing*, 27(2):431–445, 2016.

[5] N. Cordier, B. Menze, H. Delingette, and N. Ayache. Patch-based segmentation of brain tissues. 2013.

[6] L. Cui, Z. Qi, Z. Chen, F. Meng, and Y. Shi. Pavement distress detection using random decision forests. In *International Conference on Data Science*, pages 95–102. Springer, 2015.

[7] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, 36(5):961–1005, 1990.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[9] R. M. Haralick et al. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] H. Huang, R. He, Z. Sun, and T. Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1697, 2017.

[13] N. Kondo, M. Harada, and Y. Takagi. Efficient training for automatic defect classification by image augmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 226–233, March 2018.

[14] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[15] S. Z. Li. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

[16] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1925–1934, 2017.

[17] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 773–782, 2018.

[18] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li. Deepcrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing*, 338:139–153, 2019.

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[20] C. Mandriota, M. Nitti, N. Ancona, E. Stella, and A. Distante. Filter-based feature selection for rail defect detection. *Machine Vision and Applications*, 15(4):179–185, 2004.

[21] D. Mery and C. Arteta. Automatic defect recognition in x-ray testing using computer vision. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1026–1035, March 2017.

[22] B. Paniagua, M. A. Vega-Rodríguez, J. A. Gomez-Pulido, and J. M. Sanchez-Perez. Improving the industrial classification of cork stoppers by using image processing and neuro-fuzzy computing. *Journal of Intelligent Manufacturing*, 21(6):745–760, 2010.

[23] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[24] S. Qin, P. Ren, S. Kim, and R. Manduchi. Robust and accurate text stroke segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 242–250, March 2018.

[25] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[26] F. Schroff, A. Criminisi, and A. Zisserman. Object class segmentation using random forests. In *British Machine Vision Conference*, pages 1–10, 2008.

[27] L. Shang, Q. Yang, J. Wang, S. Li, and W. Lei. Detection of rail surface defects based on cnn image recognition and classification. In *2018 20th International Conference on Advanced Communication Technology (ICACT)*, pages 45–51. IEEE, 2018.

[28] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen. Automatic road crack detection using random structured forests. *IEEE Transactions on Intelligent Transportation Systems*, 17(12):3434–3445, 2016.

[29] V. A. Sindagi and S. Srivastava. Domain adaptation for automatic oled panel defect detection using adaptive support vector data description. *International Journal of Computer Vision*, 122(2):193–211, 2017.

[30] K. Y. Song, J. Kittler, and M. Petrou. Defect detection in random colour textures. *Image and vision Computing*, 14(9):667–683, 1996.

[31] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.

[32] X. Tao, D. Zhang, W. Ma, X. Liu, and D. Xu. Automatic metallic surface defect detection and recognition with convolutional neural networks. *Applied Sciences*, 8(9):1575, 2018.

[33] D.-M. Tsai and T.-Y. Huang. Automated surface inspection for statistical textures. *Image and Vision Computing*, 21(4):307–323, 2003.

[34] D. Unay and B. Gosselin. Apple defect segmentation by artificial neural networks. In *Proceedings of BeNeLux Conference on Artificial Intelligence*. Citeseer, 2006.

[35] V. V. Valindria, N. Pawlowski, M. Rajchl, I. Lavdas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 547–556, March 2018.

[36] F. M. Vilnrotter, R. Nevatia, and K. E. Price. Structural analysis of natural textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):76–89, 1986.

[37] C.-Y. Wen, S.-H. Chiu, W.-S. Hsu, and G.-H. Hsu. Defect segmentation of texture images with wavelet transform and a co-occurrence matrix. *Textile Research Journal*, 71(8):743–749, 2001.

[38] W. K. Wong, C. Yuen, D. Fan, L. Chan, and E. Fung. Stitching defect detection and classification using wavelet transform and bp neural network. *Expert Systems with Applications*, 36(2):3845–3856, 2009.

[39] X. Xie and M. Mirmehdi. Texture exemplars for defect detection on random textures. In *International Conference on Pattern Recognition and Image Analysis*, pages 404–413. Springer, 2005.

[40] X. Xie and M. Mirmehdi. Texems: Texture exemplars for defect detection on random textured surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1454–1464, 2007.

[41] N. Yu, X. Shen, Z. Lin, R. Mech, and C. Barnes. Learning to detect multiple photographic defects. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1387–1396, March 2018.