小波和原型增强查询Transformer用于像素级表面缺陷检测

冯岩[1], 蒋晓恒[1,2,3*], 陆杨[1,2,3*], 曹佳乐[4], 陈东 [1,2,3] 和 徐明良 [1,2,3][1]郑州大学[2]教育部智能集群系统工程研究中心[3]郑州国家超级计算中心 [4]天津大学ieyanfeng@163.com {jiangxiaoheng, ieylu, chendongai, iexumingliang}@zzu.edu.cn connor@tju.edu.cn

摘要

作为智能制造的重要组成部分，像素级表面缺陷检测（*SDD*）旨在通过掩码预测定位缺陷区域。先前方法采用图像无关的静态卷积来不加区分地分类每个像素特征进行掩码预测，这导致在一些具有挑战性的场景（如弱缺陷和杂乱背景）中结果次优。在本文中，受基于查询的方法的启发，我们提出了一种用于表面缺陷检测的小波和原型增强查询*Transformer*（*WP-Former*）。具体而言，通过双域*Transformer*解码器更新一组用于掩码预测的动态查询。首先，提出了一种小波增强交叉注意力（*WCA*），它在小波域中聚合图像特征的具有意义的高频和低频信息以细化查询。*WCA*通过捕获不同频率分量之间的多尺度关系来增强高频分量的表示，使查询能够更关注缺陷细节。其次，提出了一种原型引导交叉注意力（*PCA*），通过空间域中的元原型来细化查询。原型从图像特征中聚合语义上有意义的标记，帮助查询在杂乱背景下面向关键缺陷信息。在三个缺陷检测数据集（即*ESDIs-SOD*、*CrackSeg9k*和*ZJU-Leaper*）上的大量实验表明，所提出的方法在缺陷检测方面实现了*state-of-the-art*性能。代码将在 *https://github.com/yfhdm/WPFormer*.

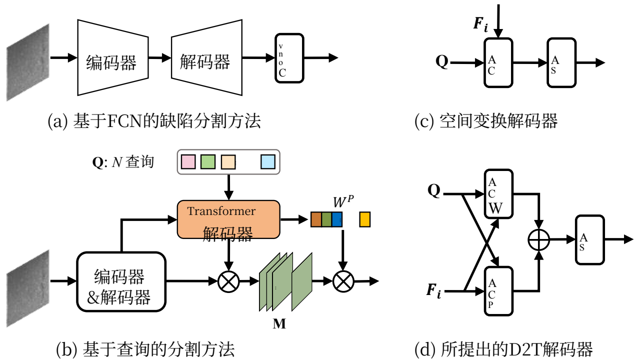


图1. (a) 之前的像素级缺陷检测方法使用静态卷积层进行掩码预测。 (b) 在基于查询的分割方法中，一组动态查询 **Q** 通过 Transformer解码器进行掩码预测。掩码预测是通过聚合具有权重 W^P 的所有查询的掩码预测 **M** 获得的。 (c) 空间域中的现有 Transformer解码器层。 (d) 所提出的双域 Transformer (D2T) 解码器层包括频率域中的小波增强交叉注意力 (WCA) 和空间域中的原型引导交叉注意力 (PCA)。

1. 引言

表面缺陷检测 (SDD) 是工业制造中的一项重要任务。自动缺陷检测提高了生产线 [2, 41] 和基础设施 [24], 的质量检验效率，能够快速识别缺陷。像素级表面缺陷检测方法旨在通过掩码预测定位缺陷区域，提供细粒度检测结果。然而，与自然场景中的目标检测不同，工业表面缺陷检测面临着诸如弱外观（特征为尺寸小、细长形状、与背景高度相似）以及复杂背景噪声等挑战。

*通讯作者：蒋晓恒，陆杨。

Wavelet and Prototype Augmented Query-based Transformer for Pixel-level Surface Defect Detection

Feng Yan[1], Xiaoheng Jiang[1,2,3*], Yang Lu[1,2,3*], Jiale Cao[4], Dong Chen [1,2,3] and Mingliang Xu [1,2,3]
[1]Zhengzhou University
[2]Engineering Research Center of Intelligent Swarm Systems, Ministry of Education
[3]National Supercomputing Center in Zhengzhou [4]Tianjin University
ieyanfeng@163.com {jiangxiaoheng, ieylu, chendongai, iexumingliang}@zzu.edu.cn
connor@tju.edu.cn

Abstract

As an important part of intelligent manufacturing, pixel-level surface defect detection (SDD) aims to locate defect areas through mask prediction. Previous methods adopt the image-independent static convolution to indiscriminately classify per-pixel features for mask prediction, which leads to suboptimal results for some challenging scenes such as weak defects and cluttered backgrounds. In this paper, inspired by query-based methods, we propose a Wavelet and Prototype Augmented Query-based Transformer (WP-Former) for surface defect detection. Specifically, a set of dynamic queries for mask prediction is updated through the dual-domain transformer decoder. Firstly, a Wavelet-enhanced Cross-Attention (WCA) is proposed, which aggregates meaningful high- and low-frequency information of image features in the wavelet domain to refine queries. WCA enhances the representation of high-frequency components by capturing multi-scale relationships between different frequency components, enabling queries to focus more on defect details. Secondly, a Prototype-guided Cross-Attention (PCA) is proposed to refine queries through meta-prototypes in the spatial domain. The prototypes aggregate semantically meaningful tokens from image features, facilitating queries to aggregate crucial defect information under the cluttered backgrounds. Extensive experiments on three defect detection datasets (i.e., ESDIs-SOD, CrackSeg9k, and ZJU-Leaper) demonstrate that the proposed method achieves state-of-the-art performance in defect detection. The code will be available at https://github.com/yfhdm/WPFormer.

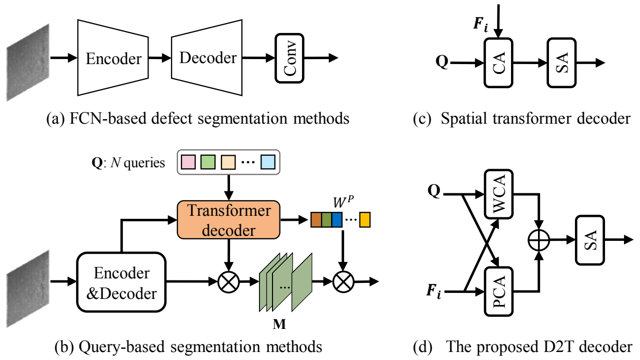

Encoder
Decoder
Conv
(a) FCN-based defect segmentation methods
F_i
Q
CA
SA
(c) Spatial transformer decoder
Q: *N* queries
Transformer decoder
W^P
Encoder &Decoder
M
(b) Query-based segmentation methods
Q
WCA
F_i
PCA
SA
(d) The proposed D2T decoder

Figure 1. (a) The previous pixel-level defect detection methods use the static convolution layer for mask prediction. (b) In the query-based segmentation methods, a set of dynamic queries **Q** are refined through a transformer decoder for mask prediction. The mask prediction is obtained by aggregating mask predictions **M** of all queries with weights W^P. (c) The existing transformer decoder layer in the spatial domain. (d) The proposed dual-domain transformer (D2T) decoder layer includes Wavelet-enhanced Cross-Attention (WCA) in the frequency domain and Prototype-guided Cross-Attention (PCA) in the spatial domain.

1. Introduction

Surface defect detection (SDD) is an important task in industrial manufacturing. Automated defect detection improves the efficiency of quality inspections across production lines [2, 41] and infrastructure [24], enabling the rapid identification of defects. Pixel-level SDD methods aim to localize defect regions with mask prediction, which provides fine-grained detection results. However, different from object detection in natural scenes, industrial surface defect detection faces challenges such as weak appearances characterized by small sizes, elongated shapes, and high similarity to the background, as well as complex background noise.

[*] Corresponding authors: Xiaoheng Jiang, Yang Lu.

针对这些挑战，已经出现了越来越多的基于深度学习的方法。这些方法遵循全卷积网络 (FCN) [27], 通过特征细化来提高网络性能 [9, 11], 多任务学习策略 [36],以及改进损失函数 [5]。尽管这些方法获得了更具判别性的特征，但在上述挑战性场景下仍存在一些不准确预测，例如遗漏缺陷细节和背景干扰的误检。我们认为一个重要原因是这些模型在预测层使用卷积层来不加区分地分类所有特征进行掩码预测。如图 1 (a) 所示，该预测过程相当于直接使用一个静态查询对视觉特征进行掩码预测。这样的查询是图像无关的，缺乏语义表示，导致缺陷检测结果次优。

近年来，基于查询的Transformer解码器架构在图像分割方面取得了令人印象深刻的结果 [4,7, 8]。这些架构引入了一个Transformer解码器，用于从图像特征中动态学习一组可学习的查询，以进行掩码预测。如图1 (b)所示，此类方法的主要目标是在查询和视觉特征之间实现语义交互，以进行掩码预测。在缺陷检测方面，这些方法存在两个主要问题。首先，这些方法 [7, 19]主要只关注空间域中的查询-特征交互。仅靠空间信息很难检测到弱缺陷。在频率域分析 [10, 44, 46], 中，高频分量和低频分量分别描述了纹理细节和物体的基本结构。高频分量包含丰富的边缘细节，这些是检测某些弱缺陷物体的关键点。因此，对于缺陷检测来说，从图像特征中聚合有意义的高频和低频分量以优化查询非常重要。其次，一些现有方法 [7, 12] 计算图像特征和查询之间的完整成对交互。冗余的背景信息可能会削弱查询对关键缺陷信息的注意力。尽管Mask2Former [8] 和PEM [4]通过掩码注意力和原型选择机制减少了空间冗余信息，但这些方法需要一个强大的掩码先验。如果先验掩码缺少某些缺陷细节或包含对背景干扰的错误预测，这些问题将被传递到后续的解码过程，导致次优结果。解决这个问题的方法是从图像特征中有效编码关键的缺陷信息，以实现高效的查询-特征交互。

为此，我们提出了一种用于表面缺陷检测的Wavelet和Prototype增强查询式Transformer，旨在增强查询与频率域和空间域中的多尺度特征之间的语义交互，首先，引入了小波增强交叉注意力（WCA），以更关注频率域中的缺陷细节。WCA利用Haar小波变换将图像特征分解为低频和高频分量。高频信息可能包含背景噪声。WCA通过捕获不同频率分量之间的全局和局部关系来生成多尺度通道权重，以调节高频分量进行噪声抑制。此外，引入了原型引导交叉注意力（PCA），以关注具有可学习元原型的关键缺陷信息，用于在空间域中更新查询。PCA动态地从图像特征中聚合原型，以关注缺陷的关键判别信息。这些原型通过捕获它们与查询之间的多尺度关系来跨通道细化查询。

总之，我们论文的主要贡献如下:

1. 我们提出了一种用于表面缺陷检测的小波和原型增强查询Transformer，该Transformer利用频率信息和空间原型来丰富用于掩码预测的查询。2. 我们提出了一种小波增强交叉注意力模块，该模块将特征中的高频和低频信息集成起来与查询交互。它可以引导查询更关注频率域中的判别性特征。3. 我们提出了一种原型引导交叉注意力模块，该模块将特征编码为有意义的原型，以在通道上细化查询。它可以减少特征冗余信息并保留用于交互的有用信息。4. 在三个公共缺陷数据集上的大量实验表明，所提出的方法在缺陷检测场景中实现了最先进的性能。

2. 相关工作

2.1. 像素级表面缺陷检测

近年来，针对工业缺陷分割提出了许多方法。这些方法的主流可以分为三种策略。1) 特征增强：这些方法引入注意力机制或上下文模块来增强或丰富特征表示。Cheng等人 [9]利用掩码预测作为指导来帮助增强特征表示。Wang等人 [37] 从不同方向捕获全局信息来细化融合特征。Cui等人 [11] 通过全局自相关增强特征上下文。Liu等人 [25]结合全局和局部注意力来学习全局和局部语义以定位缺陷区域。2) 多任务学习: [17],[36], [33] 利用边缘相关语义特征来更好地分割缺陷的边界区域。3) 改进损失

There have been more and more deep learning-based methods to solve these challenges. These methods follow the paradigm of Fully Convolutional Networks (FCN) [27], which improve the performance of networks by feature refinement [9, 11], multi-task learning strategy [36], and improving loss function [5]. Although these methods obtain more discriminative features, there are still some inaccurate predictions under the above challenging scenes, such as missing defect details and false detection of background distractions. We argue that one important reason is that these models use a convolutional layer to indiscriminately classify all features for mask prediction in the prediction layer. As shown in Fig.1 (a), this prediction process is equivalent to directly using one static query on visual features for mask prediction. Such a query is image-independent and lacks semantic representation, which leads to sub-optimal results for defect detection.

Recently, query-based transformer decoder architectures have shown impressive results in image segmentation [4, 7, 8]. These architectures introduce a transformer decoder to dynamically learn a set of learnable queries from image features for mask predictions. As shown in Fig.1 (b), the primary goal of such methods is to enable semantic interaction between queries and visual features for mask predictions. For defect detection, there are two main problems with these methods. Firstly, these methods [7, 19] mostly focus only on the query-feature interaction in the spatial domain. It is hard to detect weak defects with only spatial information. In the frequency domain analysis [10, 44, 46], high- and low-frequency components describe texture details and the basic structure of objects, respectively. The high-frequency components contain rich edge details, which are breakthrough points to detect some weak defect objects. Therefore, it is important for defect detection to aggregate meaningful high-frequency and low-frequency components from image features to refine queries. Secondly, some existing methods [7, 12] compute full pairwise interaction between image features and queries. Redundant background information may dilute the attention of queries to critical defect information. Although Mask2Former [8] and PEM [4] reduce spatial redundant information by masked attention and prototype selection mechanism, these methods require a strong mask prior. If the prior mask lacks some defect details or contains false predictions of background distractions, these problems will be transmitted to the subsequent decoding process, resulting in sub-optimal results. The solution to this problem is to effectively encode crucial defect information from image features for efficient query-feature interaction.

To this end, we propose a Wavelet and Prototype augmented query-based Transformer for surface defect detection, which aims to enhance semantic interaction between queries and multi-scale features in the frequency and spatial domains. Firstly, a Wavelet-enhanced Cross-Attention (WCA) is introduced to focus more on defect details in the frequency domain. WCA leverages the Haar wavelet transform to decompose image features into low- and high-frequency components. The high-frequency information may contain background noise. WCA generates multi-scale channel weights by capturing global and local relationships between different frequency components to modulate the high-frequency component for noise suppression. In addition, Prototype-guided Cross-Attention (PCA) is introduced to focus on crucial defect information with learnable meta-prototypes for updating queries in the spatial domain. PCA dynamically aggregates prototypes from image features to focus on crucial discriminative information about defects. These prototypes refine queries across channels by capturing multi-scale relationships between them and queries.

In summary, the main contributions of our paper are as follows:

1. We propose a Wavelet and Prototype Augmented Query-based Transformer for surface defect detection, which utilizes frequency information and spatial prototypes to enrich queries for mask prediction.
2. We present a Wavelet-enhanced Cross-Attention module, which integrates high- and low-frequency information from features to interact with queries. It can guide queries to focus more on discriminative features in the frequency domain.
3. We present a Prototype-guided Cross-Attention module, which encodes features into meaningful prototypes to refine queries over channels. It reduces the redundant information of features and retains useful information for interaction.
4. Extensive experiments on three public defect datasets demonstrate that the proposed method achieves state-of-the-art performance in defect detection scenes.

2. Related Works

2.1. Pixel-level Surface Defect Detection

Recently, many works have been proposed for industrial defect segmentation. The mainstream of these methods can be divided into three strategies. 1) Feature enhancement: These methods introduce the attention mechanism or context module to enhance or enrich feature representation. Cheng et al. [9] leveraged mask predictions as guidance to help enhance feature representations. Wang et al. [37] captured global information from different directions to refine the fused features. Cui et al. [11] enhanced feature contexts through global auto-correlation. Liu et al. [25] combined global and local attention to learn global and local semantics for locating defect regions. 2) Multi-task learning: [17], [36], [33] exploited edge-related semantic features to better segment the boundary area of defects. 3) Improving loss

函数：一些方法通过设计不同的损失函数（如自适应成本敏感损失 [22] 和聚类启发式损失 [5]）来学习更具判别性的特征。

2.2. 频率域学习

频率域分析方法已被广泛应用于计算机视觉任务，如图像分类和伪装目标检测。这些方法旨在通过小波变换和DCT等频率变换方法将RGB图像或特征转换为频率域。例如，Qin等人[32]在DCT域中提出了多光谱通道注意力，以关注更多频率分量。 [44, 45] 结合了DCT域的频率先验特征和RGB特征进行二值图像分割。Yang等人 [40]在小波域中分解高频和低频特征，以获得通道和空间注意力权重。Zhou等人 [46]通过小波变换从图像中获取高低频分量，并引入双编码器和解码器来学习和融合不同的频率分量。

2.3. 基于查询的方法

自DETR [3], 出现以来，基于查询的方法已逐渐应用于图像分割。这些方法在Transformer解码器中引入一组查询，通过查询和图像特征之间的语义关系来优化这些查询，以获得预测。MaskFormer [7] 通过掩码分类公式证明了此类方法在图像分割中的有效性。一些方法通过引入任务特定或语义查询来提高分割网络的性能。例如，Dong等人 [12] 结合了掩码和边界查询进行实例分割。He等人 [19] 引入了来自图像特征的额外查询进行分割。此外，一些方法通过增强查询和图像特征之间的语义交互来提高分割性能。例如，Cheng等人[8] 提出了掩码交叉注意力，通过应用相似度图的掩码机制，迫使查询仅关注前景特征。Cavagnero等人 [4] 提出了原型掩码交叉注意力，通过掩码交叉注意力生成原型，然后在原型和查询之间添加逐元素交互。

在本文中，我们提出了一种用于表面缺陷检测的小波和原型增强查询式Transformer，它增强了小波域和空间域的查询-特征交互。在小波域中，查询通过调制高频和低频特征进行更新，以更关注缺陷细节。在空间域中，查询通过原型进行更新，原型可以更关注关键缺陷信息。

3. 方法

3.1. 整体架构

如图2所示，所提出的WPFormer采用PVTv2作为backbone，以获得具有1/4、1/8、1/16和1/32分辨率的四级特征。所有特征的数量通过1×1 卷积调整为64个通道，并输入到基础FPN [23] 以获得1/4尺度的高分辨率特征 F_1 和从高到低分辨率的多尺度侧输出特征 $F_2 \sim F_4$。引入一组查询 $\mathbf{Q}$ $\in \mathbb{R}^{N\times D}$ 基于 F_1生成掩码预测，其中 N 和 D 分别表示每个查询的数量和通道维度。首先，$\mathbf{Q}$ 在两层Transformer F_1 内部进行更新 [8]。然后，更新的$\mathbf{Q}$ 被输入到双域Transformer (D2T) 解码器中，以通过从低到高的特征金字塔丰富查询表示 $F_2 \sim F_4$。在每个解码器块中，我们采用交叉注意力和自注意力的顺序来更新查询，遵循先前的工作 [8]。引入小波增强交叉注意力（WCA）和原型增强交叉注意力（PCA），分别在小波域和空间域聚合有意义的图像特征以更新查询。通过自注意力层，模型可以捕获查询的全局关系，以进一步丰富查询的表示。最后，每个解码器块的输出查询$\mathbf{Q}_{out}$ 被输入到基于查询的分割头进行掩码预测。

3.2. 小波增强交叉注意力

小波变换将特征分解为不同的频率分量，同时保留空间信息。在频率域中，高频分量包含丰富的边界细节，这有利于检测一些弱缺陷。然而，小波变换采用固定滤波器进行频率分解。获得的高频分量缺乏语义，可能包含噪声细节，导致错误预测。因此，有必要调制高频分量以进行噪声抑制。受此启发，我们提出了一种小波增强交叉注意力（WCA）模块，用于使用增强的频率特征细化查询。WCA的详细结构如图2 (a)所示。

给定图像特征 $F_i \in \mathbb{R}^{H_i\times W_i\times D}$，我们使用Haar小波变换将 F_i 分解为四个半分辨率特征子带： F_{LL}、 F_{LH}、 F_{HL}和 $F_{HH}$$\in \mathbb{R}^{H_i/2\times W_i/2\times D}$，其中 F_{LL} 表示低频信息 F^l自由能。 F_{LH}、 F_{HL}和 F_{HH}表示水平、垂直和对角线方向的高频细节。我们通过组合三个高频特征子带获得高频特征 F^h_{fre}。数学上，我们有：

function: Some methods learn more discriminative features by designing different loss functions such as adaptive cost-sensitive loss [22] and clustering-inspired loss [5].

2.2. Frequency Domain Learning

Frequency domain analysis methods have been widely used in computer vision tasks such as image classification and camouflaged object detection. These methods aim to transform RGB images or features into frequency domain through frequency transform methods such as wavelet transform and DCT. For example, Qin et al. [32] proposed multi-spectral channel attention in the DCT domain to focus on more frequency components. [44, 45] combined frequency prior features from the DCT domain and RGB features for binary image segmentation. Yang et al. [40] decomposed high-frequency and low-frequency features in the wavelet domain to obtain channel and spatial attention weights. Zhou et al. [46] obtained high- and low-frequency components from images with wavelet transform and introduced dual encoder and decoder to learn and fuse different frequency components.

2.3. Query-Based Methods

Since the advent of DETR [3], query-based methods have gradually been applied to image segmentation. These methods introduce a set of queries in the transformer decoder, which optimizes these queries by semantic relations between queries and image features to obtain predictions. MaskFormer [7] demonstrated the effectiveness of such methods for image segmentation with a mask classification formulation. Some methods improve the performance of the segmentation networks by introducing task-specific or semantic queries. For example, Dong et al. [12] combined mask and boundary queries for instance segmentation. He et al. [19] introduced extra queries from image features for segmentation. Moreover, some methods improve the performance of the segmentation by enhancing semantic interaction between queries and image features. For example, Cheng et al. [8] proposed masked cross-attention, which forces queries to focus only on foreground features by applying the masking mechanism for the similarity map. Cavagnero et al. [4] proposed prototype-based masked cross-attention, which generates prototypes through masked cross-attention and then adds element-wise interaction between prototypes and queries.

In this paper, we propose a Wavelet and Prototype Augmented Query-based Transformer for surface defect detection, which enhances query-feature interaction across wavelet and spatial domains. In the wavelet domain, queries are updated with modulated high-frequency and low-frequency features to focus more on defect details. In the spatial domain, queries are updated with prototypes that can focus more on crucial defect information.

3. Method

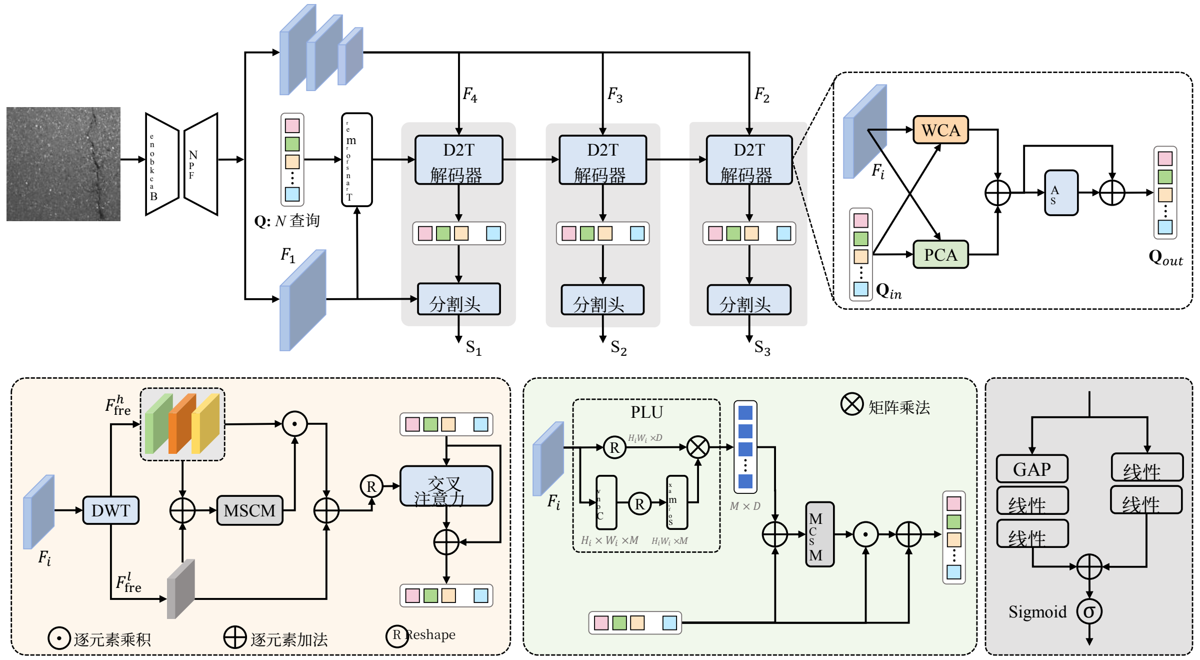
3.1. Overall Architecture

As shown in Fig.2, the proposed WPFormer adopts the PVTv2 as the backbone to obtain four-level features with 1/4, 1/8, 1/16, and 1/32 resolutions, respectively. The channel number of all features is adjusted into 64 channels by 1×1 convolution and fed into the vanilla FPN [23] to obtain 1/4 scale high-resolution features F_1 and side-output multi-scale features $F_2 \sim F_4$ from high- to low- resolutions. A set of queries $\mathbf{Q} \in \mathbb{R}^{N\times D}$ is introduced to generate mask prediction based on F_1, where N and D represent the number and the channel dimension of each query. Firstly, $\mathbf{Q}$ is updated with F_1 inside a two-layer transformer [8]. Then, the updated $\mathbf{Q}$ is fed into the Dual-Domain Transformer (D2T) decoders to enrich query representation with feature pyramid $F_2 \sim F_4$ from low- to high-resolution. In each decoder block, we adopt the order of cross- and self-attention to update queries, following the previous work [8]. Wavelet-enhanced Cross-Attention (WCA) and Prototype-enhanced Cross-Attention (PCA) are introduced to aggregate meaningful image features in the frequency and spatial domains to update queries, respectively. Through the self-attention layer, the model can capture global relationships of queries to further enrich the representation of queries. Finally, the output queries $\mathbf{Q}_{out}$ of each decoder block is fed into the query-based segmentation head for mask prediction.

3.2. Wavelet-enhanced Cross Attention

Wavelet transform decomposes features into different frequency components while preserving spatial information. In the frequency domain, high-frequency components contain rich boundary details, which are beneficial for detecting some weak defects. However, wavelet transform adopts fixed filters for frequency decomposition. The obtained high-frequency components lack semantics and may contain noise details, which lead to false predictions. So it is necessary to modulate high-frequency components for noise suppression. Inspired by this, we propose a Wavelet-enhanced Cross-Attention (WCA) module to refine queries with enhanced frequency features. The detailed structure of WCA is shown in Fig. 2 (a).

Given image features $F_i \in \mathbb{R}^{H_i\times W_i\times D}$, we use Haar wavelet transform to decompose F_i into four feature subbands with half resolutions: F_{LL}, F_{LH}, F_{HL}, and F_{HH} $\in \mathbb{R}^{H_i/2\times W_i/2\times D}$, where F_{LL} represent low-frequency information F^l_{fre}. F_{LH}, F_{HL}, and F_{HH} represent high-frequency details in the horizontal, vertical, and diagonal directions. We obtain high-frequency features F^h_{fre} by combining three high-frequency feature subbands. Mathemati-



(a) WCA (b) 主成分分析 (c) 多尺度上下文模块

图2。所提出的方法的架构。我们采用PVTv2作为backbone，并使用FPN作为像素解码器以获得多尺度特征 $\{F_i\}_{i=1}^4$。在输入Transformer解码器之前，**Q** 首先通过一个双层Transformer [8]使用高分辨率特征 F_1 进行更新。为了丰富查询的表示，我们引入双域Transformer (D2T) 解码器，通过小波增强交叉注意力 (WCA) 和原型引导交叉注意力 (PCA) 来聚合有意义的图像特征 $\{F_i\}_{i=2}^4$。查询输出和高分辨率特征被输入到分割头以生成掩码预测。

在数学上，我们有：

$$F^l_{\text{fre}} = F_{LL} \tag{1}$$

$$F^h_{\text{fre}} = F_{LH} + F_{HL} + F_{HH} \tag{2}$$

随后，我们利用不同频率分量之间的全局和局部通道依赖关系来调制高频分量。 F^h_{fre} 和 F^l自由能 被加在一起，然后输入到多尺度上下文模块 (MSCM) 以生成多尺度通道权重 $W^c_g \in \mathbb{R}^{1\times1\times D}$ 和 $W^c_l \in \mathbb{R}^{H_i/2\times W_i/2\times D}$。如图2(c)所示，全局通道权重 W^c_g 通过学习全局依赖关系来抑制特征通道的噪声，而局部通道权重 W^c_l通过学习局部依赖关系来抑制特征空间像素的噪声。在 W^c_g 和W^c_l融合后，通过sigmoid函数生成注意力权重，通过逐元素乘法来细化高频分量。数学上，我们有：

$$W^c_g = \text{Linear}(\delta(\text{Linear}(\text{GAP}(F^h_{\text{fre}} + F^l_{\text{fre}})))) \tag{3}$$

$$W^c_l = \text{Linear}(\delta(\text{Linear}(F^h_{\text{fre}} + F^l_{\text{fre}}))) \tag{4}$$

$$F^h_{\text{fre}}{}' = \sigma(W^c_g + W^c_l) \odot F^h_{\text{fre}} \tag{5}$$

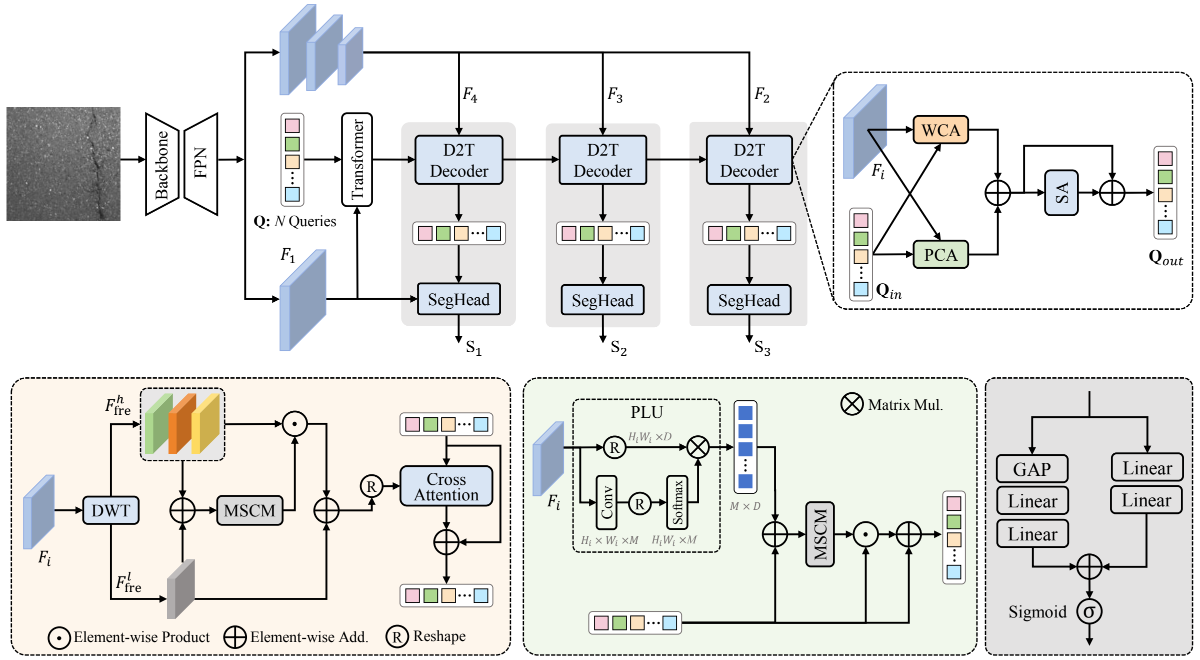
where Linear() 表示线性层。GAP 表示全局平均池化操作。 $\odot$ 表示元素-逐元素乘法。 δ 和 σ 分别表示 ReLU 和 Sigmoid 函数。

考虑到高频和低频特征对准确检测的重要性，调制的高频分量与低频特征相结合，生成特征 F'_{fre}。然后我们使用 F'_{fre} 作为键和值，通过交叉注意力层更新查询 $\mathbf{Q}_{in}$ 。从数学上讲，我们有：

$$\mathbf{Q}' = Norm(\mathbf{Q}_{in} + Attention(\mathbf{Q}_{in}, F'_{\text{fre}})) \tag{6}$$

3.3. 原型引导交叉注意力

在空间域中，标准交叉注意力 [7] 中的完全成对空间相似性会带来冗余信息。尽管 [4, 8] 中的掩码机制可以过滤掉不相关的信息并增强对缺陷区域的关注。然而，掩码注意力的有效性严重依赖于掩码先验的质量。如果掩码先验包含不完整或错误的缺陷预测，它可能会阻碍后续解码层捕捉关键细节，导致缺陷区域检测不完整。为此，我们提出了原型引导交叉注意力 (P CA) 来从两个方面减少图像特征中的冗余空间信息。首先，我们引入一个原型学习单元 (PLU) 来学习图像特征的语义簇，作为



(a) WCA (b) PCA (c) MSCM

Figure 2. The architecture of the proposed method. We adopt PVTv2 as the backbone and the FPN as the pixel decoder to obtain multi-scale features $\{F_i\}_{i=1}^4$. Before being fed into the transformer decoder, **Q** is first updated with high-resolution features F_1 by a two-layer transformer [8]. To enrich the representation of queries, we introduce Dual-Domain Transformer (D2T) decoders to aggregate meaningful image features $\{F_i\}_{i=2}^4$ via Wavelet-enhanced Cross-Attention (WCA) and Prototype-guided Cross-Attention (PCA). The query output and high-resolution features are fed into the segmentation head to generate mask prediction.

cally, we have:

$$F^l_{\text{fre}} = F_{LL} \tag{1}$$

$$F^h_{\text{fre}} = F_{LH} + F_{HL} + F_{HH} \tag{2}$$

Subsequently, we leverage global and local channel-wise dependencies between different frequency components to modulate high-frequency components. F^h_{fre} and F^l_{fre} are added together and then fed into multi-scale context module (MSCM) to generate multi-scale channel weights $W^c_g \in \mathbb{R}^{1\times1\times D}$ and $W^c_l \in \mathbb{R}^{H_i/2\times W_i/2\times D}$. As shown in Fig. 2(c), global channel weights W^c_g suppress noise from feature channels by learning global dependencies, while local channel weights W^c_l suppress noise from feature spatial pixels by learning local dependencies. After W^c_g and W^c_l are fused, attention weights are generated through the sigmoid function to refine high-frequency components via element-wise multiplication. Mathematically, we have:

$$W^c_g = \text{Linear}(\delta(\text{Linear}(\text{GAP}(F^h_{\text{fre}} + F^l_{\text{fre}})))) \tag{3}$$

$$W^c_l = \text{Linear}(\delta(\text{Linear}(F^h_{\text{fre}} + F^l_{\text{fre}}))) \tag{4}$$

$$F^h_{\text{fre}}{}' = \sigma(W^c_g + W^c_l) \odot F^h_{\text{fre}} \tag{5}$$

where Linear() represents the linear layer. GAP represents the global average pooling operation. $\odot$ represents element-wise multiplication. δ and σ represent ReLU and Sigmoid functions, respectively.

Considering the importance of both high-frequency and low-frequency features for accurate detection, the modulated high-frequency component is combined with low-frequency features, resulting in feature F'_{fre}. Then we use F'_{fre} as the key and value to update queries $\mathbf{Q}_{in}$ through the cross-attention layer. Mathematically, we have:

$$\mathbf{Q}' = Norm(\mathbf{Q}_{in} + Attention(\mathbf{Q}_{in}, F'_{\text{fre}})) \tag{6}$$

3.3. Prototype-guided Cross Attention

In the spatial domain, full pairwise spatial similarity in the standard cross-attention [7] brings redundant information. Although the masking mechanism in [4, 8] can filter out irrelevant information and enhance focus on defect regions. However, the effectiveness of masked attention heavily relies on the quality of the mask prior. If the mask prior contains incomplete or false defect prediction, it may hinder subsequent decoding layers from capturing critical details, resulting in incomplete defect region detection. To this end, we propose Prototype-guided Cross-Attention (PCA) to reduce redundant spatial information from image features from two aspects. First, we introduce a prototype learning unit (PLU) to learn semantic clusters of image features as

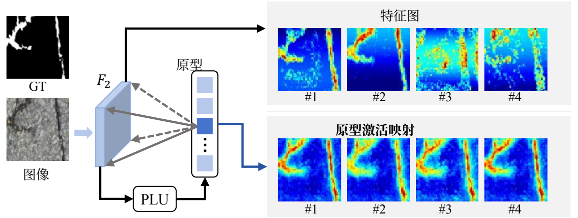


图3. 视觉比较原始特征图和原型激活特征图，针对特征 F_2。可以观察到原始特征图包含冗余的背景噪声，而所提出的原型可以更专注于关键的缺陷信息。

原型。原型自适应地聚合来自视觉特征的 informative tokens。其次，我们捕获原型和查询之间的全局和局部关系以细化查询。PCA 的详细结构如图 2(b)

原型学习单元。图像视觉特征 $F_i \in\mathbb{R}^{H_i\times W_i\times D}$ 被输入到 3×3 卷积层和 1×1卷积层，结果为 $F_i' \in \mathbb{R}^{H_i\times W_i\times M}$，其中 M 表示原型的数量，即 $M = N$。 F_i' 被展平并输入到 Softmax 函数，结果为 $F_i^w \in \mathbb{R}^{H_iW_i\times M}$。 F_i^w 的转置与展平的特征 F_i 进行矩阵乘法，得到原型特征 $F_{\text{pro}} \in \mathbb{R}^{M\times D}$。数学上，我们有：

$$F_{\text{pro}} = \text{Softmax}(F_i')^T \otimes F_i \tag{7}$$

哪里 Softmax 被应用于扁平化 F_i'的 3，如图 3所示，原型激活的特征图与原始特征图相比，更关注缺陷区域。

通过 F_{pro} 获得，$\mathbf{Q}_{in}$ 通过捕获 F_{pro} 与 $\mathbf{Q}_{in}$ 之间的多尺度关系进行细化。具体来说，F_{pro} 和 $\mathbf{Q}_{in}$ 通过元素级求和进行整合，并通过MSCM生成多尺度通道权重。多尺度通道权重使查询能够关注查询和原型之间的全局和局部空间关系。数学上，我们有：

$$W_g^c = \text{Linear}(\delta(\text{Linear}(\text{GAP}(F_{\text{pro}} + \mathbf{Q}_{in})))) \tag{8}$$
$$W_l^c = \text{Linear}(\delta(\text{Linear}((F_{\text{pro}} + \mathbf{Q}_{in})))) \tag{9}$$
$$\mathbf{Q}' = \text{Norm}(\sigma(W_g^c + W_l^c) \odot \mathbf{Q}_{in} + \mathbf{Q}_{in}) \tag{10}$$

请注意，所提出的 PCA 和 PEM-CA 之间存在两个关键差异。首先，PEM-CA 通过掩码交叉注意力获取原型，而 PCA 通过自适应聚类学习原型。其次，在查询-原型交互方面，PEM-CA 仅捕获局部关系，而 PCA 捕获全局和局部关系。

3.4. 分割头

为获得分割预测，我们使用输出查询 $\mathbf{Q}_{out}$ 对 $1/4$ 分辨率的特征图 F_1 进行解码。F_1 通过 1×1 卷积层线性投影到掩码特征 F_{mask} 中。遵循先前工作 [7, 8], $\mathbf{Q}_{out}$ 被输入到一个 3 层 MLP 中，并与展平的掩码特征的转置 F'_{mask} 相乘，以获得掩码预测 $\mathbf{M} \in \mathbb{R}^{N\times HW}$。从数学上讲，我们有

$$\mathbf{M} = \mathcal{F}_{\text{mlp}}(\mathbf{Q}_{out}) \otimes (F'_{\text{mask}})^{\text{T}} \tag{11}$$

$\mathbf{M}$ 被重塑为 $\mathbf{M}' \in \mathbb{R}^{N\times H\times W}$。然后我们利用$\mathbf{Q}_{out}$ 生成权重来融合掩码预测 $\mathbf{M}'$。权重 $W^p \in \mathbb{R}^N$ 通过线性层获得。数学上，掩码预测的生成方式如下：

$$\mathbf{S}_i = \sigma(\sum_{n=1}^{N} W_i^p \mathbf{M}'_n) \tag{12}$$

3.5. 损失函数

WPFormer采用3层Transformer解码器。在[8], 之后，我们为每个Transformer解码器添加监督。对于 *i-th* 层，输出查询被输入到分割头，基于高分辨率特征预测 S_i 。我们将掩码预测 $\{S_i\}_{i=1}^3$ 作为最终预测，即 $S_1+S_2+S_3$。此外，来自Transformer的更新查询被输入到分割头，以预测用于加性损失的掩码 S_0 。数学上，总损失函数的计算方式如下：

$$\mathcal{L}_{\text{total}} = \sum_{i=0}^{3} \mathcal{L}(S_i, G) + \mathcal{L}(S_1 + S_2 + S_3, G) \tag{13}$$

其中 G 表示真实标签。每个损失是二元交叉熵损失 ($\mathcal{L}_{BCE}$) 和交并比损失 ($\mathcal{L}_{IoU}$) 的组合，其定义为：$\mathcal{L}$ =$\mathcal{L}_{BCE} + \mathcal{L}_{IoU}$。

4. 实验

4.1. 数据集和评估指标

数据集。为了验证所提方法的有效性，我们在三个大规模缺陷数据集 ESDIs-SOD [11], CrackSeg9k[21], 和 ZJU- Leaper [42]上进行了实验。**ESDIs-SOD** 是一个综合条纹缺陷数据集。它包含14种缺陷类型，共有4800张缺陷图像，其中3600张图像用作训练集，1200张图像用作测试集。**CrackSeg9k** 是一个裂缝分割数据集。在不同类型表面上共有8051张裂缝图像，其中7243张为训练图像，395张为测试图像。**ZJU-Leaper**是一个织物缺陷数据集，包含15761张训练图像和7945张测试图像。

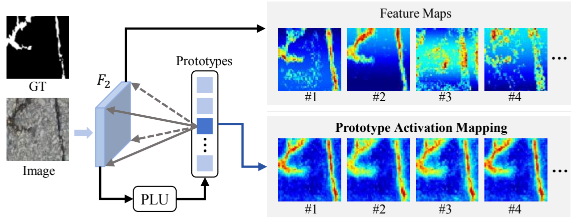


Figure 3. Visual comparison of original feature maps and prototype-activated feature maps for the features F_2. It can be observed that original feature maps contain redundant background noise and the proposed prototypes can focus more on crucial defect information.

prototypes. The prototypes adaptively aggregate informative tokens from visual features. Second, we capture global and local relationships between prototypes and queries to refine queries. The detailed structure of PCA is shown in Fig. 2 (b).

Prototype Learning Unit. The image visual features $F_i \in \mathbb{R}^{H_i\times W_i\times D}$ are fed into 3×3 convolution layer and 1×1 convolution layer, resulting $F_i' \in \mathbb{R}^{H_i\times W_i\times M}$, where M represent the number of prototypes, i.e., $M = N$. F_i' is flattened and fed into the Softmax function, resulting in $F_i^w \in \mathbb{R}^{H_iW_i\times M}$. The transpose of F_i^w performs matrix multiplication with the flattened feature F_i, resulting in prototype features $F_{\text{pro}} \in \mathbb{R}^{M\times D}$. Mathematically, we have:

$$F_{\text{pro}} = \text{Softmax}(F_i')^T \otimes F_i \tag{7}$$

where Softmax is applied on the first dimension of the flattened F_i'. As shown in Fig.3, it can be seen that prototype-activated feature maps focus more on defect regions compared with original feature maps.

With F_{pro} obtained, $\mathbf{Q}_{in}$ is refined by capturing multi-scale relationships between F_{pro} and $\mathbf{Q}_{in}$. Specifically, F_{pro} and $\mathbf{Q}_{in}$ are integrated with element-wise summation and generate multi-scale channel weights through MSCM. Multi-scale channel weights enable queries to focus on global and local spatial relationships between queries and prototypes. Mathematically, we have:

$$W_g^c = \text{Linear}(\delta(\text{Linear}(\text{GAP}(F_{\text{pro}} + \mathbf{Q}_{in})))) \tag{8}$$
$$W_l^c = \text{Linear}(\delta(\text{Linear}((F_{\text{pro}} + \mathbf{Q}_{in})))) \tag{9}$$
$$\mathbf{Q}' = \text{Norm}(\sigma(W_g^c + W_l^c) \odot \mathbf{Q}_{in} + \mathbf{Q}_{in}) \tag{10}$$

Note that there are two key differences between the proposed PCA and PEM-CA. Firstly, PEM-CA obtains prototypes by masked cross-attention, while PCA learns prototypes by adaptive clustering. Secondly, in terms of query-prototype interaction, PEM-CA captures only local relationships, whereas PCA captures both global and local relationships.

3.4. Segmentation Head

To obtain the segmentation prediction, we use output queries $\mathbf{Q}_{out}$ to decode feature maps F_1 at $1/4$ resolution. F_1 is linearly projected into mask features F_{mask} through 1×1 convolution layer. Following the previous works [7, 8], $\mathbf{Q}_{out}$ is fed into a 3-layer MLP and multiplied with the transposed of the fattened mask features F'_{mask} to obtain mask predictions $\mathbf{M} \in \mathbb{R}^{N\times HW}$. Mathematically, we have

$$\mathbf{M} = \mathcal{F}_{\text{mlp}}(\mathbf{Q}_{out}) \otimes (F'_{\text{mask}})^{\text{T}} \tag{11}$$

$\mathbf{M}$ is reshaped into $\mathbf{M}' \in \mathbb{R}^{N\times H\times W}$. Then we leverage $\mathbf{Q}_{out}$ to generate weights to fuse mask predictions $\mathbf{M}'$. The weights $W^p \in \mathbb{R}^N$ is obtained through a linear layer. Mathematically, the mask prediction is generated as follows:

$$\mathbf{S}_i = \sigma(\sum_{n=1}^{N} W_i^p \mathbf{M}'_n) \tag{12}$$

3.5. Loss Function

WPFormer adopts a 3-layer transformer decoder. Following [8], we add supervision for each transformer decoder. For the *i-th* layer, output queries are fed into the segmentation head to predict S_i based on high-resolution features. We integrate mask predictions $\{S_i\}_{i=1}^3$ as the final prediction, i.e., $S_1+S_2+S_3$. In addition, the updated queries from the transformer are fed into the segmentation head to predict mask S_0 for additive loss. Mathematically, the total loss functions are calculated as follows:

$$\mathcal{L}_{\text{total}} = \sum_{i=0}^{3} \mathcal{L}(S_i, G) + \mathcal{L}(S_1 + S_2 + S_3, G) \tag{13}$$

where G denotes the ground truth. Each loss is a combination of binary cross-entropy loss ($\mathcal{L}_{BCE}$) and Intersection over Union loss($\mathcal{L}_{IoU}$), which is defined as: $\mathcal{L} = \mathcal{L}_{BCE} + \mathcal{L}_{IoU}$.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. To validate the effectiveness of the proposed method, we conduct experiments on three large-scale defect datasets ESDIs-SOD [11], CrackSeg9k[21], and ZJU-Leaper [42]. **ESDIs-SOD** is a comprehensive strip defect dataset. It contains 14 types of defects, with a total of 4800 defect images, of which 3600 images are used as training set and 1200 images are used as test set. **CrackSeg9k** is a crack segmentation dataset. There are 8051 crack images on different types of surfaces, with 7243 training images and 395 test images. **ZJU-Leaper** is a fabric defect dataset, containing 15761 images for training and 7945 images for testing.

方法	Year	ESDIs-SOD					CrackSeg9k					ZJU-Leaper				
		$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$mF_\beta\uparrow$	$mE_\xi\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$mF_\beta\uparrow$	$mE_\xi\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$mF_\beta\uparrow$	$mE_\xi\uparrow$
JTFN [9]	ICCV’ 2021	.0188	.8730	.8975	.8723	.9560	.0166	.6849	.7947	.6977	.9043	.0251	.6898	.7801	.7194	.8742
SINetV2 [15]	TPAMI’ 2022	.0208	.8603	.8961	.8581	.9549	.0195	.6513	.7855	.6534	.8918	.0248	.7107	.8006	.7359	.8975
Mask2Former[8]	CVPR’ 2022	.0197	.8767	.9075	.8745	.9574	.0147	.7442	.8385	.7478	.9363	.0206	.7663	.8267	.7879	.9227
BBRF [28]	TIP’ 2023	.0205	.8632	.8882	.8668	.9511	.0161	.6909	.8026	.6965	.9139	.0229	.7265	.7959	.7479	.8996
PUENet [43]	TIP’ 2023	.0199	.8721	.8995	.8727	.9553	.0168	.6976	.8104	.6991	.9155	.0241	.7200	.7998	.7445	.9008
FPNet [10]	ACM MM’ 2023	.0191	.8758	.9115	.8698	.9581	.0150	.7425	.8286	.7378	.9316	.0207	.7656	.8271	.7859	.9214
MENet [39]	CVPR’ 2023	.0218	.8576	.8924	.8555	.9508	.0177	.6701	.7937	.6754	.9000	.0259	.6946	.7849	.7259	.8808
FSPNet [20]	CVPR’ 2023	.0218	.8503	.8984	.8533	.9405	.0175	.6595	.8178	.6735	.8571	.0232	.7093	.8230	.7447	.8832
FEDER [18]	CVPR’ 2023	.0219	.8553	.8922	.8530	.9505	.0189	.6604	.7865	.6633	.9028	.0269	.6890	.7795	.7166	.8905
MSCAFNet [26]	TCSVT’ 2023	.0186	.8807	.9080	.8781	.9609	.0146	.7429	.8390	.7478	.9381	.0212	.7578	.8219	.7815	.9195
A3Net [11]	TIM’ 2023	.0183	.8863	.9049	.8821	.9639	.0160	.7079	.8177	.7131	.9329	.0217	.7488	.8170	.7736	.9160
IdeNet [16]	TIP’ 2024	.0184	.8822	.9096	.8788	.9615	.0143	.7510	.8407	.7572	.9387	.0193	.7778	.8279	.8014	.9267
ZoomNeXt [30]	TPAMI’ 2024	.0195	.8754	.9047	.8717	.9581	.0150	.7371	.8286	.7409	.9329	.0192	.7803	.8317	.7994	.9282
FSEL [35]	ECCV’ 2024	.0181	.8814	.9113	.8750	.9626	.0144	.7475	.8408	.7484	.9395	.0197	.7728	.8249	.7908	.9279
CamoDiffusion [6]	AAAI’ 2024	.0188	.8809	.8948	.8767	.9614	.0163	.7239	.8228	.7150	.9274	.0259	.7167	.7974	.7377	.9006
EMCAD [34]	CVPR’ 2024	.0197	.8739	.9065	.8759	.9517	.0147	.7349	.8350	.7386	.9348	.0212	.7572	.8191	.7817	.9194
PEM[4]	CVPR’ 2024	.0198	.8747	.9102	.8725	.9557	.0146	.7414	.8333	.7452	.9354	.0208	.7632	.8233	.7852	.9202
Ours		**.0171**	**.8901**	**.9136**	**.8865**	**.9656**	**.0135**	**.7672**	**.8493**	**.7679**	**.9481**	**.0175**	**.7972**	**.8404**	**.8146**	**.9356**

表1. 各种方法在三个不同的缺陷数据集上的定量比较 结果。每个指标的最佳结果以粗体显示。

评估指标。为了定量评估各种方法的性能，我们采用了以下广泛使用的评估指标：平均绝对误差 (M) [31],平均 F-measure (mF_β, $\beta^2 = 0.3$)[1], 加权F-measure (F_β^w, $\beta^2 = 1$) [29], S度量 (S_α, $\alpha = 0.5$) [13], 平均E度量 (mE_ξ) [14], 精确率-召回率 (PR) 曲线和F-measure曲线。

4.2. 实现细节

该网络由 PyTorch 实现，并采用在 ImageNet 上预训练的 PVTv2 [38] 作为 backbone。默认情况下，我们使用 16 个可学习的查询进行掩码预测。所有实验均在 RTX 3090 GPU 上进行。我们采用 Adam 优化器，学习率为 8e-5，并使用余弦退火学习率调度器来训练网络。该网络在 ESDIs-SOD 上训练了 150 个 epoch，批大小为 8；在 ZJU-Leaper 上训练了 24 个 epoch，批大小为 4；在 Crack-Seg9k 上训练了 60 个 epoch，批大小为 4。在训练和测试阶段，输入图像被调整为 384×384 并输入网络。

4.3. 与 state-of-the-art 的比较

在本节中，我们将提出的方法与 17 种 state-of-the-art 方法进行比较，包括 JTFN [9], SINetV2 [15],Mask2Former [8], BBRF [28], PUENet [43], FPNet [10],MENet [39], FSPNet [20], FEDER [18], MSCAFNet [26],A3Net [11], IdeNet [16], ZoomNeXt [30], FSEL [35],CamoDiffusion [6], EMCAD [34],和PEM [4]。

定量比较. 表 1 列出了所提出的方法和 17 种当前最先进的方法在三个缺陷数据集上的定量比较结果，具体指标为 M, F_β^w, S_α, mF_β, 和 mE_ξ。所提出的方法在现有 SDD 模型（JTFN[9]和 A3Net[11]）上表现优异。例如，与 A3Net [11], 相比，所提出的方法在 M, F_β^w, S_α, mF_β, 和 mE_ξ 指标上分别平均提升了 13.85%，5.09%，2.56%，4.49%，和 1.32%。与采用 PVTv2 作为主干网络的检测模型（MSCAFNet[26],、ZoomNeXt[30], 、IdeNet[16]）相比，所提出的方法也取得了更好的性能。例如，与 IdeNet[16], 相比，所提出模型在 M, F_β^w, S_α, mF_β, 和 mE_ξ 指标上分别平均提升了 7.33%，1.85%，0.99%，1.31%，和 0.80%。所提出的方法还优于现有的基于查询的分割方法，如 Mask2Former[8] 和 PEM [4], , 这些方法也采用 PVTv2 作为主干网络。与 Mask2Former[8], 相比，所提出模型在三个数据集上分别平均提升了 M, F_β^w, S_α, mF_β, 和 mE_ξ 12.14%，2.88%，1.21%，2.48%，和 1.17%。此外，图 4 展示了每个数据集上不同方法的 PR 和 F- measure 曲线。我们的 F- measure 曲线在大多数阈值下都取得了比其他方法更好的性能。

可视化比较. 图 5显示了某些方法在三个缺陷数据集上的检测结果。如第1行和第4行所示，由于缺陷与背景之间的高度相似性，一些方法难以检测完整的缺陷区域。研究发现，一些方法将某些背景干扰检测为缺陷区域，例如第2行、第3行和第5行。此外，对于一些方法来说，检测薄裂纹也具有挑战性，例如第3行和第6行。相比之下，提出的方法获得了更准确的预测。

Methods	Year	ESDIs-SOD					CrackSeg9k					ZJU-Leaper				
		$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$mF_\beta\uparrow$	$mE_\xi\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$mF_\beta\uparrow$	$mE_\xi\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$mF_\beta\uparrow$	$mE_\xi\uparrow$
JTFN [9]	ICCV’2021	.0188	.8730	.8975	.8723	.9560	.0166	.6849	.7947	.6977	.9043	.0251	.6898	.7801	.7194	.8742
SINetV2 [15]	TPAMI’2022	.0208	.8603	.8961	.8581	.9549	.0195	.6513	.7855	.6534	.8918	.0248	.7107	.8006	.7359	.8975
Mask2Former[8]	CVPR’2022	.0197	.8767	.9075	.8745	.9574	.0147	.7442	.8385	.7478	.9363	.0206	.7663	.8267	.7879	.9227
BBRF [28]	TIP’2023	.0205	.8632	.8882	.8668	.9511	.0161	.6909	.8026	.6965	.9139	.0229	.7265	.7959	.7479	.8996
PUENet [43]	TIP’2023	.0199	.8721	.8995	.8727	.9553	.0168	.6976	.8104	.6991	.9155	.0241	.7200	.7998	.7445	.9008
FPNet [10]	ACM MM’2023	.0191	.8758	.9115	.8698	.9581	.0150	.7425	.8286	.7378	.9316	.0207	.7656	.8271	.7859	.9214
MENet [39]	CVPR’2023	.0218	.8576	.8924	.8555	.9508	.0177	.6701	.7937	.6754	.9000	.0259	.6946	.7849	.7259	.8808
FSPNet [20]	CVPR’2023	.0218	.8503	.8984	.8533	.9405	.0175	.6595	.8178	.6735	.8571	.0232	.7093	.8230	.7447	.8832
FEDER [18]	CVPR’2023	.0219	.8553	.8922	.8530	.9505	.0189	.6604	.7865	.6633	.9028	.0269	.6890	.7795	.7166	.8905
MSCAFNet [26]	TCSVT’2023	.0186	.8807	.9080	.8781	.9609	.0146	.7429	.8390	.7478	.9381	.0212	.7578	.8219	.7815	.9195
A3Net [11]	TIM’2023	.0183	.8863	.9049	.8821	.9639	.0160	.7079	.8177	.7131	.9329	.0217	.7488	.8170	.7736	.9160
IdeNet [16]	TIP’2024	.0184	.8822	.9096	.8788	.9615	.0143	.7510	.8407	.7572	.9387	.0193	.7778	.8279	.8014	.9267
ZoomNeXt [30]	TPAMI’2024	.0195	.8754	.9047	.8717	.9581	.0150	.7371	.8286	.7409	.9329	.0192	.7803	.8317	.7994	.9282
FSEL [35]	ECCV’2024	.0181	.8814	.9113	.8750	.9626	.0144	.7475	.8408	.7484	.9395	.0197	.7728	.8249	.7908	.9279
CamoDiffusion [6]	AAAI’2024	.0188	.8809	.8948	.8767	.9614	.0163	.7239	.8228	.7150	.9274	.0259	.7167	.7974	.7377	.9006
EMCAD [34]	CVPR’2024	.0197	.8739	.9065	.8759	.9517	.0147	.7349	.8350	.7386	.9348	.0212	.7572	.8191	.7817	.9194
PEM[4]	CVPR’2024	.0198	.8747	.9102	.8725	.9557	.0146	.7414	.8333	.7452	.9354	.0208	.7632	.8233	.7852	.9202
Ours		**.0171**	**.8901**	**.9136**	**.8865**	**.9656**	**.0135**	**.7672**	**.8493**	**.7679**	**.9481**	**.0175**	**.7972**	**.8404**	**.8146**	**.9356**

Table 1. Quantitative comparison results of various methods on three different defect datasets. The best result for each metric is in bold.

Evaluation Metrics. To quantitatively evaluate the performance of various methods, we adopt the following widely-used evaluation metrics: Mean Absolute Error (M) [31], mean F-measure (mF_β, $\beta^2 = 0.3$)[1], weighted F-measure (F_β^w, $\beta^2 = 1$) [29], S-measure (S_α, $\alpha = 0.5$) [13], mean E-measure (mE_ξ) [14], Precision-Recall (PR) curve and F-measure curve.

4.2. Implementation Details

The network is implemented by PyTorch and adopts the PVTv2 [38] pre-trained on the ImageNet as the backbone. By default, we use 16 learnable queries for mask prediction. All experiments are conducted on an RTX 3090 GPU. We adopt Adam optimizer with a learning rate of 8e-5 and a cosine decay learning rate scheduler to train the network. The network is trained for 150 epochs with a batch size of 8 on ESDIs-SOD, for 24 epochs with a batch size of 4 on ZJU-Leaper, and 60 epochs with a batch size of 4 on Crack-Seg9k, respectively. In the training and test stage, the input image is resized to 384 × 384 and fed into the network.

4.3. Comparisons with State-of-the-art

In this section, we compare the proposed method with 17 state-of-the-art methods, including JTFN [9], SINetV2 [15], Mask2Former [8], BBRF [28], PUENet [43], FPNet [10], MENet [39], FSPNet [20], FEDER [18], MSCAFNet [26], A3Net [11], IdeNet [16], ZoomNeXt [30], FSEL [35], CamoDiffusion [6], EMCAD [34], and PEM [4].

Quantitative comparisons. Table 1 lists quantitative comparison results for the proposed method and 17 state-of-the-art methods on three defect datasets in terms of M, F_β^w, S_α, mF_β, and mE_ξ. The proposed method outperforms existing SDD models (JTFN[9] and A3Net[11]) well. For example, compared with A3Net [11], the proposed method yields average improvements with 13.85%, 5.09%, 2.56%, 4.49%, and 1.32% in terms of M, F_β^w, S_α, mF_β, and mE_ξ, respectively. Compared with detection models that adopt PVTv2 as the backbone (MSCAFNet[26], ZoomNeXt[30], IdeNet[16]), the proposed method also achieves better performance. For example, compared with IdeNet[16], the proposed model obtains achieves average performance gains with 7.33%, 1.85%, 0.99%, 1.31%, and 0.80% in terms of M, F_β^w, S_α, mF_β, and mE_ξ, respectively. The proposed methods also outperform the existing query-based segmentation methods such as Mask2Former [8] and PEM [4], which also adopt PVTv2 as the backbone. Compared with Mask2Former[8], the proposed model improves the M, F_β^w, S_α, mF_β, and mE_ξ by 12.14%, 2.88%, 1.21%, 2.48%, and 1.17% on average across three datasets, respectively. In addition, Fig. 4 shows the PR and F-measure curves of different methods on each dataset. Our F-measure curve achieves better performance than other methods at most thresholds.

Visualization comparisons. Fig. 5 shows the detection results of some methods over three defect datasets. As shown in the 1st and 4th rows, some methods struggle to detect complete defect regions because of the high similarity between defects and backgrounds. It is found that some methods detect some background distractions as the defect areas, such as the 2nd, 3rd, and 5th rows. It is also challenging for some methods to detect thin cracks, such as the 3rd and 6th rows. By contrast, the proposed method obtains more accu-

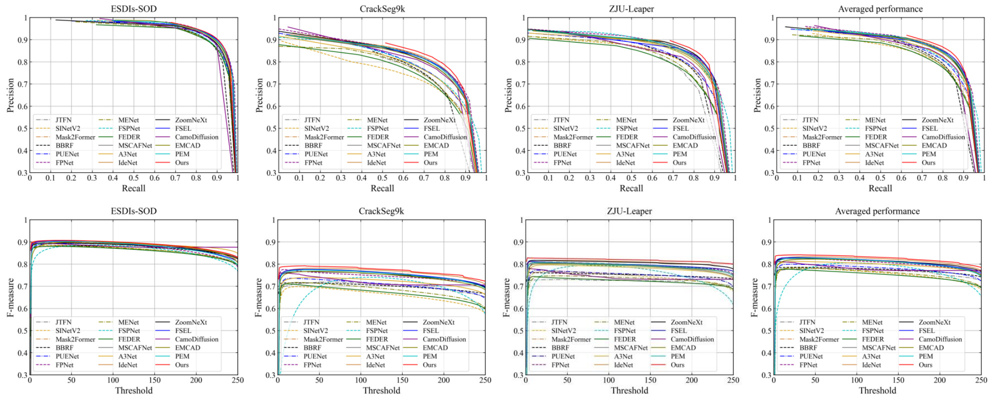


图4. 不同方法在PR和F-measure曲线方面的性能比较。

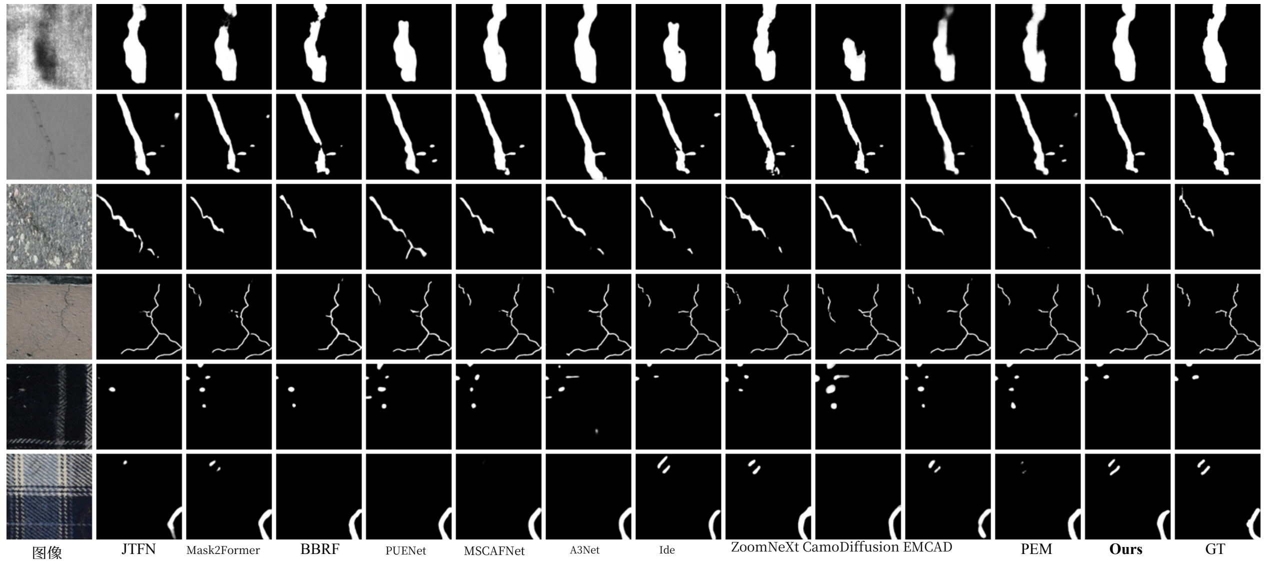


图5. 一些代表性方法的视觉比较。

率预测。

4.4. 消融研究

为了展示网络中每个组件的有效性，我们在ESDIs-SODandCrackSeg9k上分别进行消融研究。

不同的交叉注意力. 表 2 比较了Transformer解码器中不同交叉注意力模块的性能，包括传统交叉注意力、掩码交叉注意力、PEM-CA以及提出的WCA和PCA模块。掩码交叉注意力和PEM-CA中的掩码机制在一定程度上提高了模型对缺陷区域的关注，但也导致次优性能，因为先验掩码预测可能会丢失重要的缺陷细节，从而阻止查询学习这些微妙特征。可以看出，提出的双域交叉注意力从频率域和主成分分析域自适应地选择有意义的特征，这可以减少细节信息的丢失。与传统交叉注意力、掩码交叉注意力和PEM-CA相比，提出的交叉注意力在 {v3}<style id='4'>. 为了更好地展示双域交叉注意力模块的有效性，我们在图 <style id='6'>6<style id='8'>中比较了不同交叉注意力的检测结果。可以看出，传统交叉注意力、掩码交叉注意力和PEM-CA存在不完整（用红色框标记）或误检（用绿色框标记）的问题。相反，提出的交叉注意力可以更关注缺陷细节并实现准确的检测结果。

主成分分析域和空间原型视角，这可以减少细节信息的丢失。与传统交叉注意力、掩码交叉注意力和PEM-CA相比，提出的交叉注意力在 F^w_β. 为了更好地展示双域交叉注意力模块的有效性，我们在图 6中比较了不同交叉注意力的检测结果。可以看出，传统交叉注意力、掩码交叉注意力和PEM-CA存在不完整（用红色框标记）或误检（用绿色框标记）的问题。相反，提出的交叉注意力可以更关注缺陷细节并实现准确的检测结果。

查询数量. 表3 (a) 分析了查询数量对模型性能的影响。可以观察到使用少量查询可以带来显著的性能提升。当 N_q 为 16 时，模型达到最佳性能。

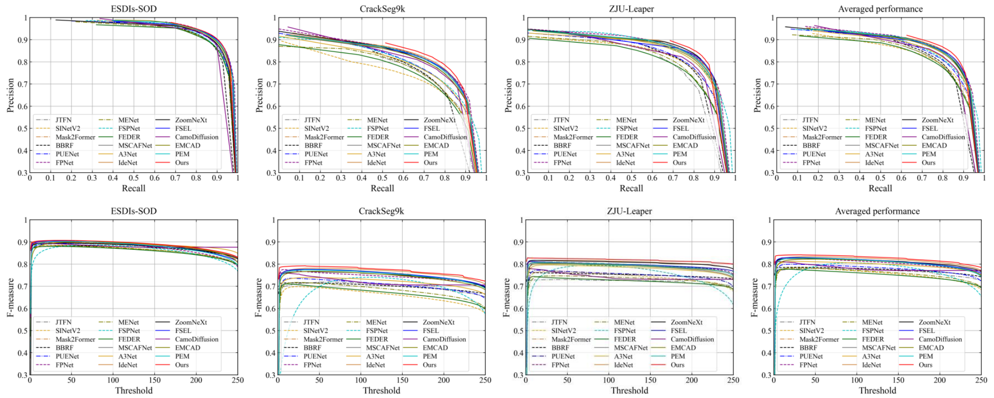


Figure 4. Performance comparisons of different methods in terms of PR and F-measure curves.

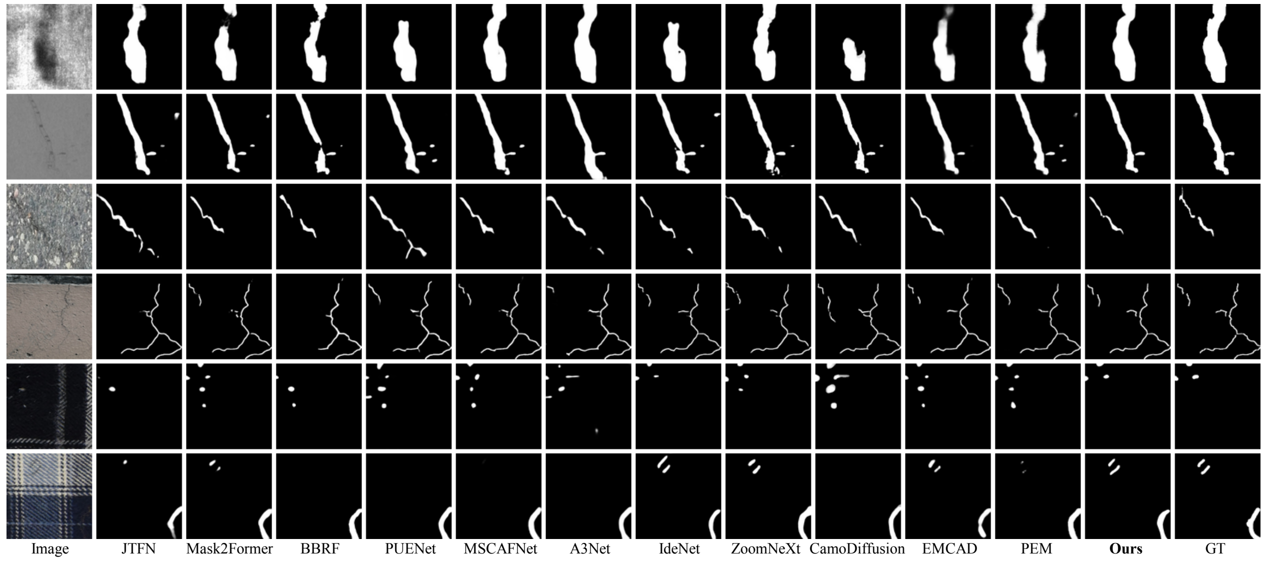


Figure 5. Visual comparisons of some representative methods.

rate predictions.

4.4. Ablation Studies

To demonstrate the effectiveness of each component presented in the network, we perform the ablation study on ESDIs-SOD and CrackSeg9k, respectively.

Different cross attention. Table 2 compares the performance of different cross-attention modules in the transformer decoder, including conventional CA, Masked CA, PEM-CA, and the proposed WCA and PCA modules. The masking mechanism in the Masked CA and PEM-CA improves the model's focus on defect areas to some extent but also leads to suboptimal performance because the prior mask prediction may lose important defect details, thereby preventing queries from learning these subtle features. It can be seen that the proposed dual domain cross-attention adaptively selects meaningful features from frequency domain and spatial prototype perspectives, which can reduce the loss of detailed information. Compared with standard CA, masked CA, and PEM-CA, the proposed cross-attention obtains averaged gains of 2.14%, 1.78%, and 1.62% in terms of F^w_β. To better demonstrate the effectiveness of the dual domain cross-attention module, we compare detection results of different cross-attention in Fig 6. It can be seen that CA, masked CA, and PEM-CA suffer from incomplete (marked by red boxes) or false detection (marked by green boxes). On the contrary, the proposed cross-attention can focus more on defect details and achieve accurate detection results.

Number of queries. Table 3 (a) analyzes the effect of the number of queries on the performance of the model. It can be observed that using a small number of queries can bring significant performance gains. The model achieves the optimal performance when N_q is 16.

方法	ESDIs-SOD			CrackSeg9k		
	$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$
w/ CA	.0193	.8778	.9090	.0146	.7458	.8361
w/ Masked-CA	.0190	.8797	.9097	.0141	.7494	.8368
w/ PEM-CA	.0187	.8802	.9077	.0142	.7513	.8400
我们的 (w/ WCA)	.0175	.8858	.9125	.0140	.7583	.8425
我们的 (w/ PCA)	.0179	.8855	.9118	.0139	.7579	.8420
我们的（两者）	**.0171**	**.8901**	**.9136**	**.0135**	**.7672**	**.8493**

表2。Transformer解码器中不同交叉注意力模块的消融实验。

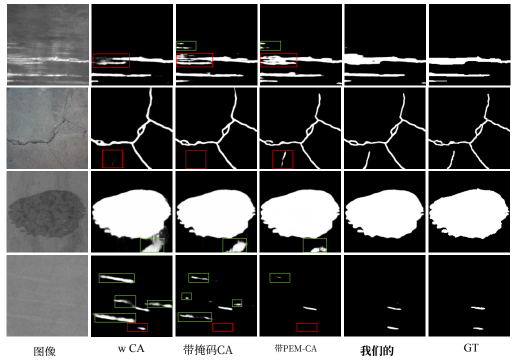


图6。表2中不同交叉注意力下检测结果的视觉比较2。红色和绿色框分别表示漏检和误检缺陷区域。

设置		ESDIs-SOD			CrackSeg9k		
	N_q	$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$
(a)	8	.0179	.8853	.9112	.0139	.7609	.8444
	16	**.0171**	**.8901**	**.9136**	**.0135**	**.7672**	**.8493**
	64	**.0171**	.8885	.9131	.0137	.7654	.8467
(b)	WCA	MAE $\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	MAE $\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$
	加	.0175	.8865	.9102	.0140	.7598	.8446
	w/ 调制	**.0171**	**.8901**	**.9136**	**.0135**	**.7672**	**.8493**
(c)	PCA	MAE $\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	MAE $\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$
	全局	.0174	.8872	.9120	.0138	.7613	.8443
	局部	.0175	.8871	.9111	.0139	.7595	.8443
	Both	**.0171**	**.8901**	**.9136**	**.0135**	**.7672**	**.8493**

表3. 对所提出网络的不同设置进行的消融研究：(a) 显示了不同查询数量的影响，(b) 显示了WCA模块内不同频率融合方法的影响，以及 (c) 显示了PCA模块内查询-原型交互模式的影响。

WCA 中不同频率成分的融合方法效果. 表 3 (b) 分析了 WCA 中不同频率融合策略的影响，即加性融合和我们的调制融合。与加性融合相比融合，自适应调制融合实现了更好的性能，在 F_β^w 上平均增益为0.69%。从图 7 可以看出，调制融合可以更专注于缺陷区域并抑制背景噪声。

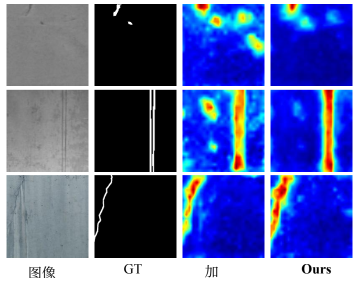


图7. 小波域中特征图 F_2 的可视化比较：具有加性融合和调制融合的频域特征。

PCA内交互模式的影响。表 3(c) 显示了PCA内查询和原型之间多尺度关系对性能的影响。全局和局部关系的组合比单尺度关系实现了更好的性能。这表明全局和局部关系都很重要，整合这些关系可以更好地更新查询。

5. 结论

在本文中，我们提出了一种用于表面缺陷检测的小波和原型增强查询式Transformer（WPFormer）。该方法使查询和特征在频率域和空间域中的交互成为可能。在小波域中，WCA使查询能够聚合精细的频率分量，从而提高其对弱缺陷细节的敏感性。在空间域中，PCA通过图像特征的动态原型丰富了查询的表示。这些原型自适应地引导查询专注于关键的缺陷区域，从而促进对缺陷区域的更精确分割。在三个缺陷检测数据集（即ESDIs-SOD、CrackSeg9k和ZJU-Leaper）上的大量实验表明，该方法在缺陷检测方面达到了state-of-the-art的性能。

6. 致谢

这项工作部分得到了中国国家自然科学基金（资助项目62172371、资助项目U21B2037、资助项目62102370）的支持；部分得到了河南省自然科学基金（资助项目232300421093）的支持；以及部分得到了河南省高校科技创新人才（资助项目25HASTIT033）的支持。

Method	ESDIs-SOD			CrackSeg9k		
	$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$
w/ CA	.0193	.8778	.9090	.0146	.7458	.8361
w/ Masked-CA	.0190	.8797	.9097	.0141	.7494	.8368
w/ PEM-CA	.0187	.8802	.9077	.0142	.7513	.8400
Ours (w/ WCA)	.0175	.8858	.9125	.0140	.7583	.8425
Ours (w/ PCA)	.0179	.8855	.9118	.0139	.7579	.8420
Ours (Both)	**.0171**	**.8901**	**.9136**	**.0135**	**.7672**	**.8493**

Table 2. Ablation for different cross-attention modules in the transformer decoder.

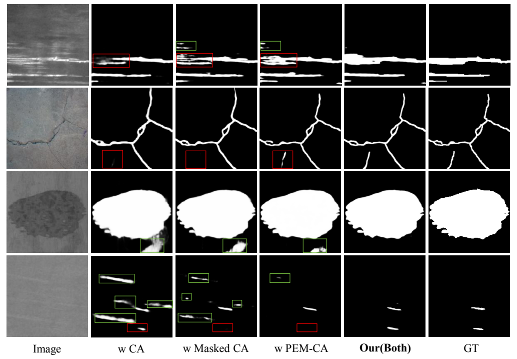


Figure 6. Visual comparison of detection results obtained with different cross-attention in Table 2. The red and green boxes denote missed and false defect areas, respectively.

Settings		ESDIs-SOD			CrackSeg9k		
	N_q	$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$
(a)	8	.0179	.8853	.9112	.0139	.7609	.8444
	16	**.0171**	**.8901**	**.9136**	**.0135**	**.7672**	**.8493**
	64	**.0171**	.8885	.9131	.0137	.7654	.8467
(b)	WCA	MAE $\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	MAE $\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$
	w/ Add	.0175	.8865	.9102	.0140	.7598	.8446
	w/ Modulation	**.0171**	**.8901**	**.9136**	**.0135**	**.7672**	**.8493**
(c)	PCA	MAE $\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	MAE $\downarrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$
	Global	.0174	.8872	.9120	.0138	.7613	.8443
	Local	.0175	.8871	.9111	.0139	.7595	.8443
	Both	**.0171**	**.8901**	**.9136**	**.0135**	**.7672**	**.8493**

Table 3. Ablation study on different settings of the proposed network: (a) shows the effect of different query numbers, (b) shows the effect of different frequency fusion methods within WCA module, and (c) shows the effect of query-prototype interaction modes within PCA module.

Effect of fusion methods for different frequency components within WCA. Table 3 (b) analyzes the effect of different frequency fusion strategies within WCA, i.e., additive fusion and our modulated fusion. Compared with additive fusion, the adaptive modulated fusion achieves better performance, with averaged gains of 0.69% in terms of F_β^w. It can be seen from Fig.7 that the modulated fusion can focus more on defect regions and suppress background noise.

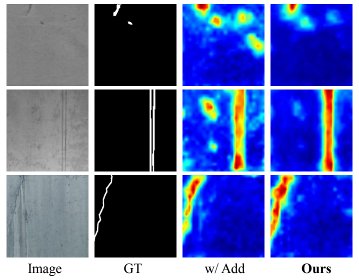


Figure 7. Visual comparison of feature maps F_2 in the wavelet domain: frequency-domain features with additive fusion and modulation fusion.

Effect of interaction modes within PCA. Table 3 (c) shows the effect of the multi-scale relationship between queries and prototypes within PCA for performance. The combinations of global and local relationships achieve better performance than single-scale relationships. This indicates that global and local relationships are both important and integrating these relationships can better update the query.

5. Conclusion

In this paper, we propose a Wavelet and Prototype Augmented Query-based Transformer (WPFormer) for surface defect detection. The proposed method enables the interaction between queries and features in both frequency and spatial domains. In the wavelet domain, WCA enables queries to aggregate refined frequency components, which enhances their sensitivity to weak defect details. In the spatial domain, PCA enriches the representation of queries through dynamic prototypes from image features. These prototypes adaptively guide the queries to focus on crucial defect regions, facilitating more accurate segmentation of defect areas. Extensive experiments on three defect detection datasets (i.e., ESDIs-SOD, CrackSeg9k, and ZJU-Leaper) demonstrate that the proposed method achieves state-of-the-art performance in defect detection.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62172371, Grant U21B2037, Grant 62102370; in part by Natural Science Foundation of Henan Province under Grant 232300421093; and in part by Scientific and Technological Innovation Talents in Universities of Henan Province 25HASTIT033.

参考文献

[1] 拉达克里希纳·阿查尔塔、希拉·赫马米、弗朗西斯科·埃斯特拉达和萨比娜·苏斯特伦克。频率调谐的显著区域检测。在IEEE/CVF计算机视觉与模式识别会议，第1597–1604页，2009年。6[2] 安-克里斯汀·贝特、帕特里克·布鲁斯、加博尔·巴拉萨斯、马蒂亚斯·卢德维希和阿洛伊斯·克诺尔。集成电路逆向工程中的自动缺陷检测。在IEEE/CVF计算机视觉应用会议，第1596–1605页，2022年。1[3] 尼古拉斯·卡里翁、弗朗西斯科·马萨、加布里埃尔·辛纳维、尼古拉斯·乌苏尼尔、亚历山大·基里洛夫和谢尔盖·扎戈鲁伊科。基于transformer的端到端目标检测。在欧洲计算机视觉会议，第213–229页，2020年。3[4]尼古洛·卡瓦内罗、加布里埃莱·罗西、克劳迪娅·库塔诺、弗朗塞斯卡·皮斯蒂利、马科·奇科内、朱塞佩·阿维塔和法比奥·塞梅利。PEM：原型高效掩码former用于图像分割。在IEEE/CVF计算机视觉与模式识别会议，第15804 15813页，2024年2、3、4、6 – 。[5] 庄庄·陈、卓楠·莱、杰·陈和建强·李。思维边缘非裂缝区域：基于聚类启发的表示学习用于裂缝分割。在IEEE/CVF计算机视觉与模式识别会议，第12698–12708页，2024年。2、3[6] 钟西·陈、科·孙和先明·林。CamoDiffusion：通过条件扩散模型进行伪装目标检测。在AAAI人工智能会议，第1272–1280页，2024年。6[7] 伯恩·程、亚历克斯·施文格和亚历山大·基里洛夫。像素级分类并非语义分割的全部。在神经信息处理系统进展，第17864–17875页，2021年。2、3、4、5[8] 伯恩·程、伊什安·米斯拉、亚历山大·G·施文格、亚历山大·基里洛夫和罗hit·吉德哈尔。掩码注意力掩码transformer用于通用图像分割。在IEEE/CVF计算机视觉与模式识别会议，第1290–1299页，2022年。2、3、4、5、6[9] 明飞·程、凯丽·赵、许红·郭、亚静·徐和俊·郭。联合拓扑保持和特征细化网络用于曲线结构分割。在IEEE/CVF国际计算机视觉会议，第7147–7156页，2021年。2、6[10] 润民·丛、梦瑶·孙、三艺·张、晓飞·周、伟·张和瑶·赵。频率感知网络用于伪装目标检测。在ACM国际多媒体会议，第1179–1189页，2023年。2、6[11] 崔文琪、宋克辰、冯虎、贾修健、刘少宁和闫云辉。自相关感知聚合网络用于带钢表面缺陷的显著目标检测。IEEE仪器仪表与测量 transactions，72:1–12，2023年。2、5、6[12] 博·董、嘉伦·裴、荣荣·高、天柱·向、硕·王和欢·熊。统一基于查询的范式用于伪装实例分割。在ACM国际多媒体会议，第2131–2138页，2023年。2、3

[13] 范登平, 程明明, 刘云, 李涛, 和 阿里·博吉. 结构度量：评估前景图的新方法. 在IEEE/CVF国际计算机视觉会议, 第4548–4557页, 2017. 6[14] 范登平, 龚程, 曹阳, 任波, 程明明, 和 阿里·博吉. 增强对齐度量用于二值前景图评估. 在国际人工智能联合会议, 第698–704页, 2018. 6[15] 范登平, 贾格鹏, 程明明, 和 邵凌. 隐蔽物体检测. IEEE模式分析与机器智能汇刊, 44(10):6024–6042, 2021. 6[16] 关久威, 方晓林, 朱通欣, 蔡志鹏, 凌振, 杨明, 和 罗军舟. IdeNet：让神经网络像生物一样识别伪装物体. IEEE图像处理汇刊, 33:4824–4839, 2024. 6[17] 韩程军, 李公阳, 和 刘智. 两阶段边缘重用网络用于条钢表面缺陷的显著目标检测. IEEE仪器仪表与测量汇刊, 71:1–12, 2022. 2[18] 何春明, 李凯, 张亚超, 唐龙祥, 张宇伦, 郭振华, 和 李秀. 基于特征分解和边缘重建的伪装物体检测. 在IEEE/CVF计算机视觉与模式识别会议, 第22046–22055页, 2023. 6[19] 何俊杰, 李鹏宇, 耿一峰, 和 谢宣松. Fastinst：一个简单的基于查询的模型用于实时实例分割. 在IEEE/CVF计算机视觉与模式识别会议, 第23663–23672页, 2023. 2, 3[20] 黄周, 戴杭, 天柱·向, 硕·王, 陈怀新, 秦杰, 和 欢·熊. 特征收缩金字塔用于基于Transformer的伪装物体检测. 在IEEE/CVF计算机视觉与模式识别会议, 第5557–5566页, 2023. 6[21]Shreyas Kulkarni, Shreyas Singh, Balakrishnan Dhananjay, Siddharth Sharma, Saipraneeth Devunuri, 和 Sai Chowdeswara Rao Korlapati. CrackSeg9k：裂缝分割数据集和框架的集合与基准. 在欧洲计算机视觉研讨会, 第179–195页, 2022. 5[22] 李凯, 王波, 田英杰, 和 齐志权. 基于自适应成本敏感损失函数的快速和准确的路面裂缝检测. 控制论IEEE汇刊, 53(2): 1051–1062, 2021. 3[23] 林宗毅, Piotr Doll´ ar, Ross Girshick, 何凯明, Bharath Hariharan, 和 Serge Belongie. 特征金字塔网络用于目标检测. 在IEEE/CVF计算机视觉与模式识别会议, 第2117– 2125页, 2017. 3[24] 刘华军, 苗祥宇, Christoph Mertz, 徐程中, 和 孔辉. Crackformer：用于细粒度裂缝检测的Transformer网络. 在IEEE/CVF国际计算机视觉会议, 页面 3783–3792, 2021. 1[25] 刘泰恒, 何兆水, 林志杰, 曹广中, 苏文庆, 和 谢胜利. 用于表面缺陷检测的自适应图像分割网络. 神经网络与学习系统IEEE汇刊, 35(6): 8510–8523, 2024. 2[13] 刘云, 李海航, 程娟, 和 陈训. Mscaf-net：通过学习多尺度上下文感知特征进行伪装物体检测的通用框架. 电路与系统视频技术IEEE汇刊, 33(9): 4934–4947, 2023. 6[14] Jonathan Long, Evan Shelhamer, 和 Trevor Darrell. 全卷积网络用于语义分割. 在IEEE/CVF计算机视觉与模式识别会议, 第3431–3440页, 2015. 2[15] 马明灿, 夏长群, 谢晨曦, 陈晓武, 和 李嘉. 增强更广泛的感受野用于显著目标检测. IEEE图像处理汇刊, 32:1026–1038, 2023. 6[16] Margolin Ran, Zelnik-Manor Lihi, 和 Tal Ayellet. 如何评估前景图. 在IEEE/CVF计算机视觉与模式识别会议, 页面 248 255, 2014 6 – . [17] 庞伟, 赵晓琪, 天柱·向, 张利和, 和 陆胡川. ZoomNext：用于伪装物体检测的统一协同金字塔网络. 模式分析与机器智能IEEE汇刊, 46 (12):9205–9220, 2024. 6[18] Federico Perazzi, Philipp Kr[19] ahenb [20] uhl, Yael Pritch, 和 Alexander Hornung. 显著性滤波器：基于对比度的滤波器用于显著区域检测. 在IEEE/CVF计算机视觉与模式识别会议, 页面 733 740, 2012 6 – . [21]秦泽群, 张鹏毅, 吴飞, 和 李希. Fcanet：频率通道注意力网络. 在IEEE/CVF国际计算机视觉会议, 页面 783 792, 2021 3 – . [22] 邱远, 刘红丽, 刘建伟, 石波, 和 李岩夫. 区域和边缘感知网络用于铁路表面缺陷分割. IEEE仪器仪表与测量汇刊, 73:1–13, 2024. 2[23] Md Mostafijur Rahman, Mustafa Munir, 和 Radu Marculescu. EmCad：用于医学图像分割的高效多尺度卷积注意力解码. 在IEEE/CVF计算机视觉与模式识别会议, 页面 11769–11779, 2024. 6´孙艳光, 许春艳, 杨健, 薛汉宇, 和 罗雷. 频率空间纠缠学习用于伪装物体检测. 在欧洲计算机视觉会议, 页面 343–360, 2024. 6[24] 万斌, 晓飞·周, 郑波伦, 韩冰, 朱尊杰, 王宏魁, 孙瑶琪, 张继勇, 和 颜成刚. Lfrnet：用于表面缺陷显著目标检测的定位、聚焦和细化网络. IEEE仪器仪表与测量汇刊, 72:1– 12, 2023. 2[25] 王楚涵, 陈海永, 和 赵慎森 . Rern：用于多晶硅太阳能电池缺陷分割的丰富边缘特征细化检测网络. 工业信息IEEE汇刊 , 20(2):1408–1419, 2023. 2

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009. 6

[2] Ann-Christin Bette, Patrick Brus, Gabor Balazs, Matthias Ludwig, and Alois Knoll. Automated defect inspection in reverse engineering of integrated circuits. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1596–1605, 2022. 1

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020. 3

[4] Niccolò Cavagnero, Gabriele Rosi, Claudia Cuttano, Francesca Pistilli, Marco Ciccone, Giuseppe Averta, and Fabio Cermelli. Pem: Prototype-based efficient maskformer for image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15804–15813, 2024. 2, 3, 4, 6

[5] Zhuangzhuang Chen, Zhuonan Lai, Jie Chen, and Jianqiang Li. Mind marginal non-crack regions: Clustering-inspired representation learning for crack segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12698–12708, 2024. 2, 3

[6] Zhongxi Chen, Ke Sun, and Xianming Lin. Camodiffusion: Camouflaged object detection via conditional diffusion models. In *AAAI Conference on Artificial Intelligence*, pages 1272–1280, 2024. 6

[7] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 17864–17875, 2021. 2, 3, 4, 5

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 2, 3, 4, 5, 6

[9] Mingfei Cheng, Kaili Zhao, Xuhong Guo, Yajing Xu, and Jun Guo. Joint topology-preserving and feature-refinement network for curvilinear structure segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 7147–7156, 2021. 2, 6

[10] Runmin Cong, Mengyao Sun, Sanyi Zhang, Xiaofei Zhou, Wei Zhang, and Yao Zhao. Frequency perception network for camouflaged object detection. In *ACM International Conference on Multimedia*, pages 1179–1189, 2023. 2, 6

[11] Wenqi Cui, Kechen Song, Hu Feng, Xiujian Jia, Shaoning Liu, and Yunhui Yan. Autocorrelation-aware aggregation network for salient object detection of strip steel surface defects. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023. 2, 5, 6

[12] Bo Dong, Jialun Pei, Rongrong Gao, Tian-Zhu Xiang, Shuo Wang, and Huan Xiong. A unified query-based paradigm for camouflaged instance segmentation. In *ACM International Conference on Multimedia*, pages 2131–2138, 2023. 2, 3

[13] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *IEEE/CVF International Conference on Computer Vision*, pages 4548–4557, 2017. 6

[14] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *International Joint Conference on Artificial Intelligence*, pages 698–704, 2018. 6

[15] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, 2021. 6

[16] Juwei Guan, Xiaolin Fang, Tongxin Zhu, Zhipeng Cai, Zhen Ling, Ming Yang, and Junzhou Luo. Idenet: Making neural network identify camouflaged objects like creatures. *IEEE Transactions on Image Processing*, 33:4824–4839, 2024. 6

[17] Chengjun Han, Gongyang Li, and Zhi Liu. Two-stage edge reuse network for salient object detection of strip steel surface defects. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022. 2

[18] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22046–22055, 2023. 6

[19] Junjie He, Pengyu Li, Yifeng Geng, and Xuansong Xie. Fastinst: A simple query-based model for real-time instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23663–23672, 2023. 2, 3

[20] Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5557–5566, 2023. 6

[21] Shreyas Kulkarni, Shreyas Singh, Dhananjay Balakrishnan, Siddharth Sharma, Saipraneeth Devunuri, and Sai Chowdeswara Rao Korlapati. Crackseg9k: a collection and benchmark for crack segmentation datasets and frameworks. In *European Conference on Computer Vision Workshops*, pages 179–195, 2022. 5

[22] Kai Li, Bo Wang, Yingjie Tian, and Zhiquan Qi. Fast and accurate road crack detection based on adaptive cost-sensitive loss function. *IEEE Transactions on Cybernetics*, 53(2): 1051–1062, 2021. 3

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 3

[24] Huajun Liu, Xiangyu Miao, Christoph Mertz, Chengzhong Xu, and Hui Kong. Crackformer: Transformer network for fine-grained crack detection. In *IEEE/CVF International Conference on Computer Vision*, pages 3783–3792, 2021. 1

[25] Taiheng Liu, Zhaoshui He, Zhijie Lin, Guang-Zhong Cao, Wenqing Su, and Shengli Xie. An adaptive image segmen-

分割网络. IEEE Trans- actions on Neural Networks and Learning Systems, 35(6): 8510–8523, 2024. 2[26] 刘云, 李海航, 程娟, 和 陈训. Mscaf-net：一个通用框架用于通过学习多尺度上下文感知特征进行伪装物体检测. 电路与系统视频技术IEEE汇刊, 33(9): 4934–4947, 2023. 6[27] Jonathan Long, Evan Shelhamer, 和 Trevor Darrell. 全卷积网络用于语义分割. 在IEEE/CVF计算机视觉与模式识别会议, 页面 3431–3440, 2015. 2[28] 马明灿, 夏长群, 谢晨曦, 陈晓武, 和 李嘉. 增强更广泛的感受野用于显著目标检测. IEEE图像处理汇刊, 32:1026– 1038, 2023. 6[29] Margolin Ran, Zelnik-Manor Lihi, 和 Tal Ayellet. 如何评估前景图. 在IEEE/CVF计算机视觉与模式识别会议, 页面 248 255, 2014 6 – . [30] 庞伟, 赵晓琪, 天柱·向, 张利和, 和 陆胡川. ZoomNext：一个统一协同金字塔网络用于伪装物体检测. 模式分析与机器智能IEEE汇刊, 46 (12):9205–9220, 2024. 6[31] Federico Perazzi, Philipp Kr¨ahenb¨uhl, Yael Pritch, 和 Alexander Hornung. 显著性滤波器：基于对比度的滤波器用于显著区域检测. 在IEEE/CVF计算机视觉与模式识别会议, 页面 733 740, 2012 6 – . [32] 秦泽群, 张鹏毅, 吴飞, 和 李希. Fcanet：频率通道注意力网络. 在IEEE/CVF国际计算机视觉会议, 页面 783 792, 2021 3 – . [33] 邱远, 刘红丽, 刘建伟, 石波, 和 李岩夫. 区域和边缘感知网络用于铁路表面缺陷分割. IEEE仪器仪表与测量汇刊, 73:1–13, 2024. 2[34] Md Mostafijur Rahman, Mustafa Munir, 和 Radu Marculescu. EmCad：高效多尺度卷积注意力解码用于医学图像分割. 在IEEE/CVF计算机视觉与模式识别会议, 页面 11769–11779, 2024. 6[35] 孙艳光, 许春艳, 杨健, 薛汉宇, 和 罗雷. 频率空间纠缠学习用于伪装物体检测. 在欧洲计算机视觉会议, 页面 343–360, 2024. 6[36] 万斌, 晓飞·周, 郑波伦, 韩冰, 朱尊杰, 王宏魁, 孙瑶琪, 张继勇, 和 颜成刚. Lfrnet：用于表面缺陷显著目标检测的定位、聚焦和细化网络. IEEE仪器仪表与测量汇刊, 72:1– 12, 2023. 2[37] 王楚涵, 陈海永, 和 赵慎森. Rern：用于多晶硅太阳能电池缺陷分割的丰富边缘特征细化检测网络. 工业信息IEEE汇刊 , 20(2):1408–1419, 2023. 2

[38] 王文海, 谢恩泽, 李翔, 范登平, 宋凯涛, 梁丁, 陆通, 罗平, 和 邵凌. Pvt v2: 基于金字塔视觉Transformer的改进基线. 计算视觉媒体, 8(3):415–424, 2022. 6[39] 王怡, 王如丽, 范欣, 王天柱 , 和 何翔健 . 像素、区域和对象：显著目标检测的多重增强. 在IEEE/CVF计算机视觉与模式识别会议, 页面 10031– 10040, 2023. 6[40] 杨宇婷, 焦立成, 刘旭, 刘方, 杨舒媛, 李玲玲, 陈不华, 李秀方, 和 黄中坚. 双小波注意力网络用于图像分类. IEEE电路与系统视频技术汇刊, 33(4):1899–1910, 2022. 3[41] 杨元福和孙敏. 基于混合经典-量子深度学习的半导体缺陷检测. 在IEEE/CVF计算机视觉与模式识别会议, 页面 2323–2332, 2022. 1[42] 张晨凯, 冯少哲, 王旭隆, 和 王月明. Zju-leaper: 用于织物缺陷检测的基准数据集及比较研究. IEEE人工智能汇刊, 1(3):219–232, 2020. 5[43] 张怡, 张静, 瓦西姆·哈米杜切, 和 奥利维耶·德福尔热. 伪装目标检测的预测不确定性估计. IEEE图像处理汇刊, 32:3580–3591, 2023. 6[44] 钟一杰, 李博, 唐吕, 库广云, 吴双, 和 丁寿红. 频率域中的伪装目标检测. 在IEEE/CVF计算机视觉与模式识别会议, 页面 4504–4513, 2022. 2, 3[45] 周岩, 博·董, 吴元峰, 朱文涛, 陈耿, 和 张延宁. 基于频先验的二值图像分割. 在国际人工智能联合会议, 页面 1822–1830, 2023. 3[46] 周岩峰, 黄嘉兴, 王陈龙, 宋乐, 和 杨歌. Xnet: 用于生物医学图像全监督和半监督语义分割的小波低频和高频融合网络. 在IEEE/CVF国际计算机视觉会议, 页面 21085–21096, 2023. 2, 3

tation network for surface defect detection. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6): 8510–8523, 2024. 2

[26] Yu Liu, Haihang Li, Juan Cheng, and Xun Chen. Mscaf-net: A general framework for camouflaged object detection via learning multi-scale context-aware features. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9): 4934–4947, 2023. 6

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 2

[28] Mingcan Ma, Changqun Xia, Chenxi Xie, Xiaowu Chen, and Jia Li. Boosting broader receptive fields for salient object detection. *IEEE Transactions on Image Processing*, 32:1026–1038, 2023. 6

[29] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2014. 6

[30] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (12):9205–9220, 2024. 6

[31] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 733–740, 2012. 6

[32] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *IEEE/CVF International Conference on Computer Vision*, pages 783–792, 2021. 3

[33] Yuan Qiu, Hongli Liu, Jianwei Liu, Bo Shi, and Yanfu Li. Region and edge-aware network for rail surface defect segmentation. *IEEE Transactions on Instrumentation and Measurement*, 73:1–13, 2024. 2

[34] Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11769–11779, 2024. 6

[35] Yanguang Sun, Chunyan Xu, Jian Yang, Hanyu Xuan, and Lei Luo. Frequency-spatial entanglement learning for camouflaged object detection. In *European Conference on Computer Vision*, pages 343–360, 2024. 6

[36] Bin Wan, Xiaofei Zhou, Bolun Zheng, Haibing Yin, Zunjie Zhu, Hongkui Wang, Yaoqi Sun, Jiyong Zhang, and Chenggang Yan. Lfrnet: Localizing, focus, and refinement network for salient object detection of surface defects. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023. 2

[37] Chuhan Wang, Haiyong Chen, and Shenshen Zhao. Rern: Rich edge features refinement detection network for polycrystalline solar cell defect segmentation. *IEEE Transactions on Industrial Informatics* , 20(2):1408–1419, 2023. 2

[38] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 6

[39] Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10031–10040, 2023. 6

[40] Yuting Yang, Licheng Jiao, Xu Liu, Fang Liu, Shuyuan Yang, Lingling Li, Puhua Chen, Xiufang Li, and Zhongjian Huang. Dual wavelet attention networks for image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4):1899–1910, 2022. 3

[41] Yuan-Fu Yang and Min Sun. Semiconductor defect detection by hybrid classical-quantum deep learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2323–2332, 2022. 1

[42] Chenkai Zhang, Shaozhe Feng, Xulongqi Wang, and Yueming Wang. Zju-leaper: A benchmark dataset for fabric defect detection and a comparative study. *IEEE Transactions on Artificial Intelligence*, 1(3):219–232, 2020. 5

[43] Yi Zhang, Jing Zhang, Wassim Hamidouche, and Olivier Deforges. Predictive uncertainty estimation for camouflaged object detection. *IEEE Transactions on Image Processing*, 32:3580–3591, 2023. 6

[44] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4504–4513, 2022. 2, 3

[45] Yan Zhou, Bo Dong, Yuanfeng Wu, Wentao Zhu, Geng Chen, and Yanning Zhang. Dichotomous image segmentation with frequency priors. In *International Joint Conference on Artificial Intelligence*, pages 1822–1830, 2023. 3

[46] Yanfeng Zhou, Jiaxing Huang, Chenlong Wang, Le Song, and Ge Yang. Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images. In *IEEE/CVF International Conference on Computer Vision*, pages 21085–21096, 2023. 2, 3