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 (WPFormer) 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.

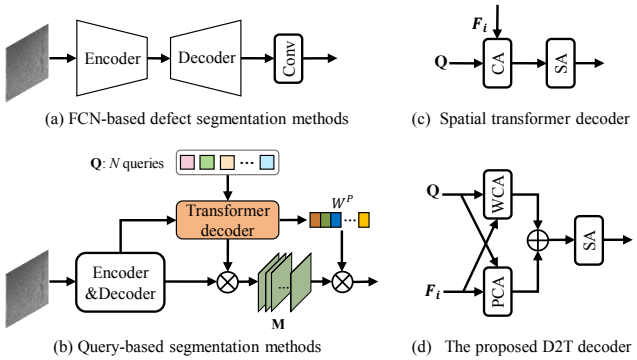


Figure 1. (a) The previous pixel-level defect detection methods use the static convolution layer for mask prediction. (b) In the query-based segmentation methods, a set of dynamic queries **Q** are refined through a transformer decoder for mask prediction. The mask prediction is obtained by aggregating 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.

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

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

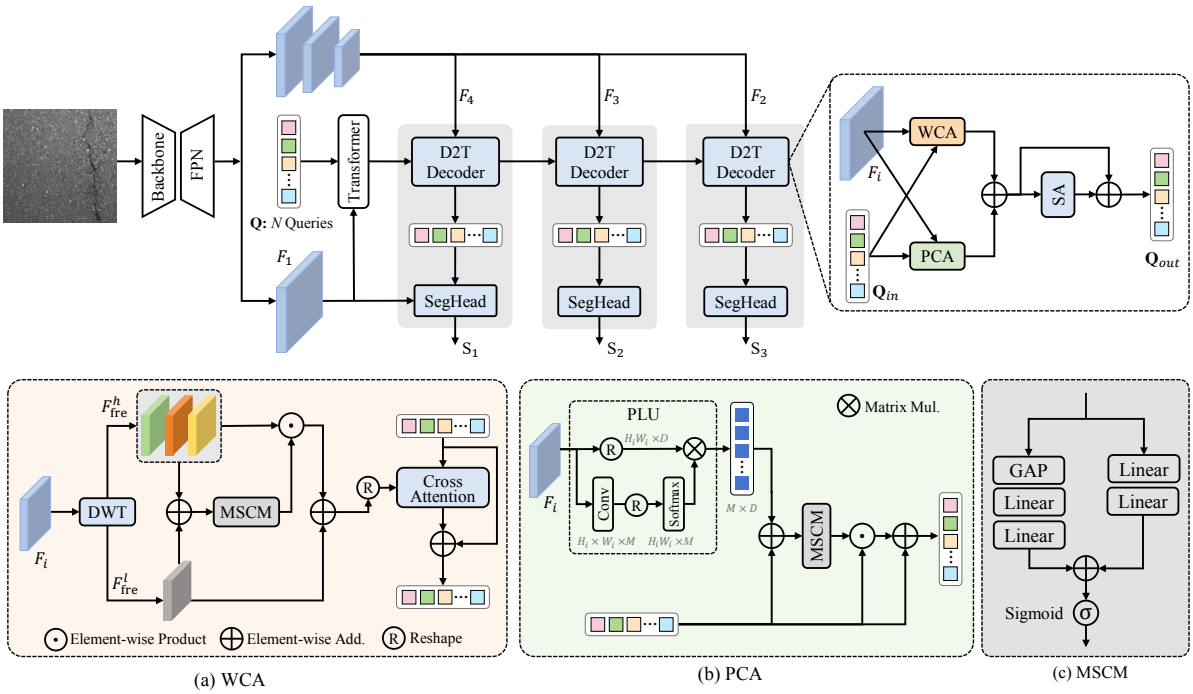
3.1. Overall Architecture

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) 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, $\mathbf{Q}$ 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:

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

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

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

$$W_g^c = \text{Linear}(\delta(\text{Linear}(\text{GAP}(F_{\text{fre}}^{h} + F_{\text{fre}}^{l})))) \tag{3}$$

$$W_l^c = \text{Linear}(\delta(\text{Linear}(F_{\text{fre}}^{h} + F_{\text{fre}}^{l}))) \tag{4}$$

$$F_{\text{fre}}^{h\,\prime} = \sigma(W_g^c + W_l^c) \odot F_{\text{fre}}^{h} \tag{5}$$

where $\text{Linear}()$ 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

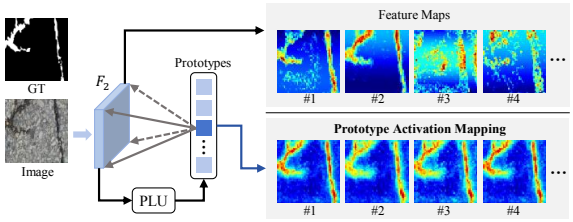


Figure 3. Visual comparison of original feature maps and prototype-activated feature maps for the features F_2. It can be observed that original feature maps contain redundant background noise and the proposed prototypes can focus more on crucial defect information.

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 $F_i^w \in \mathbb{R}^{H_i W_i \times M}$. 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.

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 CrackSeg9k, 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-

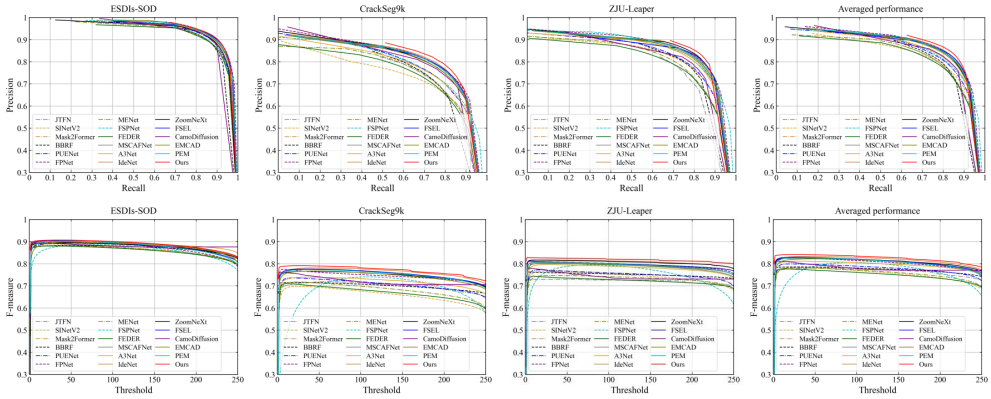


Figure 4. Performance comparisons of different methods in terms of PR and F-measure curves.

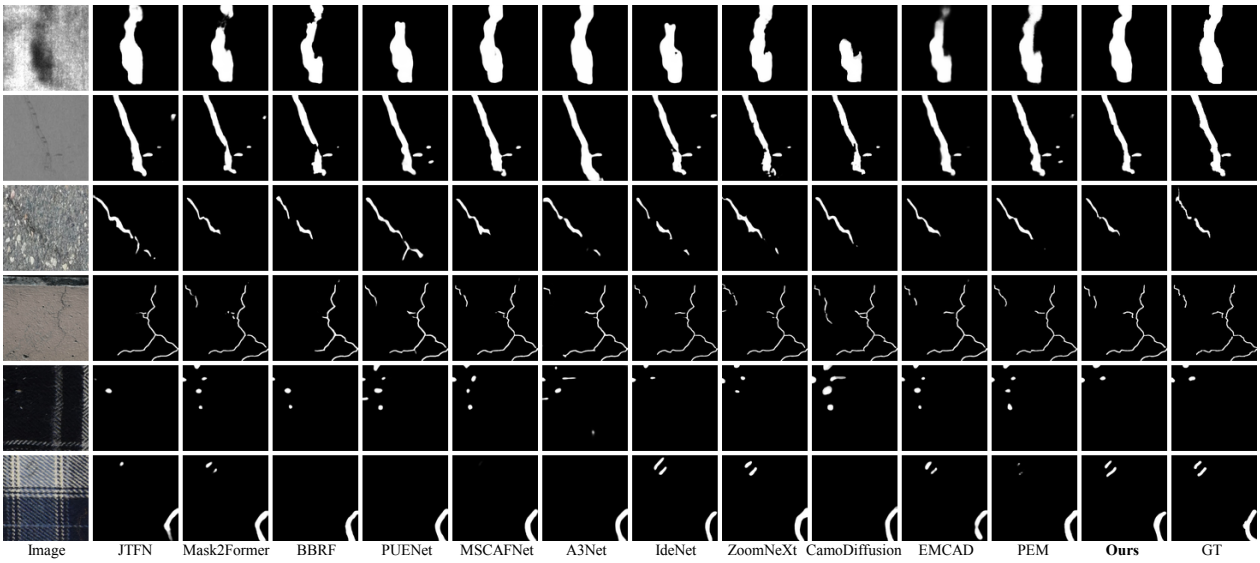


Figure 5. Visual comparisons of some representative methods.

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_β^w. 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.

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.

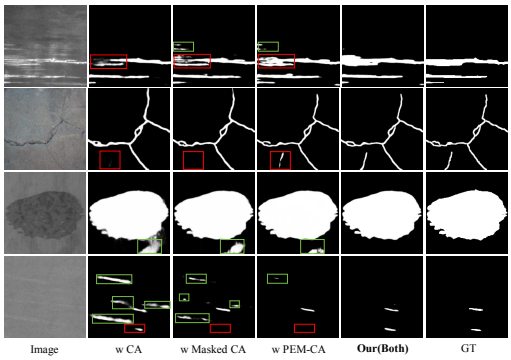


Figure 6. Visual comparison of detection results obtained with different cross-attention in Table 2. The red and green boxes denote missed and false defect areas, respectively.

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
	WCA	MAE $\downarrow$	$F_\beta^w \uparrow$	$S_\alpha \uparrow$	MAE $\downarrow$	$F_\beta^w \uparrow$	$S_\alpha \uparrow$
(b)	w/ Add	.0175	.8865	.9102	.0140	.7598	.8446
	w/ Modulation	**.0171**	**.8901**	**.9136**	**.0135**	**.7672**	**.8493**
	PCA	MAE $\downarrow$	$F_\beta^w \uparrow$	$S_\alpha \uparrow$	MAE $\downarrow$	$F_\beta^w \uparrow$	$S_\alpha \uparrow$
(c)	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.

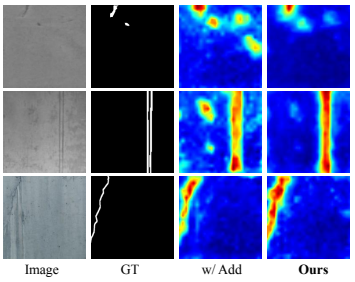


Figure 7. Visual comparison of feature maps F_2 in the wavelet domain: frequency-domain features with additive fusion and modulation fusion.

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.

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-

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