CAT-EDNet: Cross-Attention Transformer based Encoder-Decoder Network for Salient Defect Detection of Strip Steel Surface

Qiwu Luo, *Member, IEEE*, Jiaojiao Su, Chunhua Yang, *Fellow*, *IEEE*,
Weihua Gui, *Member, IEEE*,Olli Silven, *Senior Member, IEEE*, Li Liu, *Senior Member, IEEE*

***Abstract*—The morphologies of various surface defects on strip steel suffer from oil stain, water drops, steel textures and erratic illumination. It is still challenging to recognize defect boundary precisely from cluttered backgrounds. This paper emphasizes such a fact that skip connections between encoder and decoder are not equally effective, attempts to adaptively allocate the aggregation weights which represent differentiated information entropy values in channel-wise, by importing a stack of cross-attention transformer (CAT) into the encoder-decoder network (EDNet). Besides, a cross- attention refinement module (CARM) is constructed closely after the decoder to further optimize the coarse saliency maps. This newly nominated CAT-EDNet can well address the semantic gap issue among the multi-scale features for its multi-head attention structure. The CAT-EDNet performs best on insuring defect integrity and maintaining defect boundary details when compared with *twelve* state-of-the-arts, and the detection efficiency is at 28 fps even under the noise interfered scenario.**

***Index Terms*—Automated visual inspection (AVI), steel strip, salient detection, encoder-decoder network, transformer.**

I. Introduction

Strip steel is one of the fundamental products in iron and steel industry, which is widely applied in machinery, automobile manufacturing, construction, shipbuilding and even daily-used electrical products. Due to the influence of production process and rolling environment, there will inevitably be some defects on the surface of the finished strip, such as cracks, patches, scratches [1], which directly decrease the quality of the end product. Therefore, rapid and accurate surface defect detection is the primary task of strip steel quality inspection. However, the traditional manual visual sampling inspection, based on prior knowledge and probability to estimate the comprehensive quality of strip steel, has long been unable to meet the needs of modern industrial production [2]. The magnetic-flux-leakage- or eddy-current-based techniques are also suffering with the large equipment volume, low detection rate and low inspection efficiency [3]. With the development of deep learning, image-based methods can realize high accuracy and efficiency in defect inspection, and gradually become the mainstream measures. Visual attention can quickly and accurately allocate limited processing resources to prominent visual areas. Salient detection based on the above visual attention mechanism can capture the most significant and attention-attracting object in the scene image, which can achieve effective separation of foreground object and background [4]. Therefore, it has been widely used as a preprocessing operation in the tasks of defect segmentation [5], defect classification [6], defect recognition [7].

The traditional salient object detection methods [8][9] essentially depend on carefully-designed handcrafted features, objective functions and optimization strategy, which generally results in less robust and unreliable performance in complicated background. Deep learning models have received considerable critical attention for its remarkable performance on various benchmarks. The early patch-wise deep models [10] independently classify the pixels based on local features within each patch, are incompetent to achieve spatial accuracy. Many multi-level context-based architectures are also designed for salient object detection. The stacked cross refinement network [11] simultaneously refine multi-level object-aware and edge-aware features. Liu *et al*. [41] aggregate the global and local information by introducing a pyramid pooling module. BASNet [27] configures two sequentially U-like structures for boundary-aware salient object detection. Benefiting from the richer multi-level contextual features, the performances of the mentioned methods are significantly improved. However, some models introduce negative features leads to misleading inference. By embedding attention mechanism, the context selection based methods selectively integrate the effective multi-layer features. The local and global pixel-wise contextual attention is recurrently captured to predict salient maps in [40]. Innovatively, the CFPN [12] learns a set-of layer-specific weights for the effective feature selection, according to the direct cross-layer communication. In addition, some interesting approaches have also merged. The two-level nested U^2-Net [13] is powerful in extracting intra-stage multi-scale features without degrading the map resolution. The background matting technique [14] can also be transformed for salient object detection.

This work was supported jointly by the National Natural Science Foundation of China under Grant 61973323 and Grant 6201101509, by the Hunan Provincial Natural Science Foundation under Grant 2021JJ20078, and by the Science and Technology Innovation Program of Hunan Province under Grant 2021RC3019 and Grant 2021RC1001. (Corresponding Author: Chunhua Yang, Email: ychh@csu.edu.cn)

Qiwu Luo, Jiaojiao Su, and Chunhua Yang are with the School of Automation, Central South University, Changsha 410083, China.

Olli Silven is with the Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, 90014 Oulu, Finland.

Li Liu is with the College of System Engineering, National University of Defense Technology, Changsha 410073, China, and also with the Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, 90014 Oulu, Finland.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2022.3165270, IEEE Transactions on Instrumentation and Measurement

Notably in the field of salient surface defect detection for strip steel, a deeply supervised encoder-decoder residual network (EDRNet) [15] is reported to being superior to the many currently prestigious methods on both detection efficiency and noise robustness. However, for some hard samples with low contrast background, EDRNet still has limitation that some boundaries of saliency maps present un-smoothness and inaccuracy. Stimulated by this situation, we make attempt to study the underlying reasons behind it and find out improving breakthrough point to further improve the EDRNet. And we found the principal reason why the defect segmentation performance of the EDRNet decreased when facing challenging samples is that the skip connections between encoder and decoder have been set with the same weights but their descriptive abilities are always not identical (more details refer to the Section II.A). Then we propose a cross-attention transformer-based encoder-decoder network (i.e., CAT-EDNet) for salient defect detection of strip steel surface. The main contributions are as follows.

First, ***for defect integrity***, a cross-attention transformer (CAT) is embedded into the encoder-decoder network (EDNet) to dynamically allocate the aggregation weights of multi-scale layers to determine the salient region. By achieving cross-layer communication through multi-head attention structure, the salient low-level features at shallow layers are ascribed bigger weights to restore spatial structure, while high-level features in deep layers are reserved to abstractly describe the whole object.

Second, ***for defect boundary precision***, a cross-attention refinement module (CARM) is constructed closely after the decoder to further optimize the coarse saliency maps. By explicitly modelling the correlation between temporal features through CA-based residual U-block, the comprehensive prediction features are effectively focused at each fusion stage.

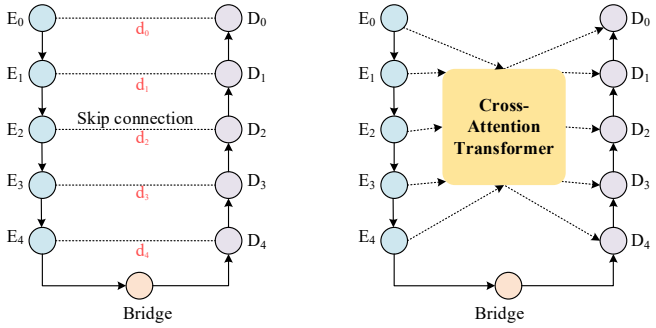
With the above cascade scheme constructed by the global-oriented CAT and the local-oriented CARM, the newly nominated CAT-EDNet can well address the semantic gap issue among the multi-scale features for its multi-head attention structure. When compared with *twelve* state-of-the-arts on challenging strip steel benchmark dataset SD-saliency-900 [15], our approach performs visually superior in defect integrity and boundary precision, shows competitive quantitative results of 93.51SM and 90.31 IoU, even at 28 fps under the severe background disturbances.

The rest of the paper is organized as follows. Section II includes the detailed motivation and related work. Section III elaborates the proposed CAT-EDNet framework. Extensive experiments and some discussions are presented in Section IV. Finally, Section V concludes this paper.

II. Motivation And Preliminaries

A. Motivation

As is shown in Fig. 2, we find the skip connections in EDRNet (Fig. 1 (a)), which helps recover the full spatial resolution through encoding-decoding process, are not equally effective. The "all" connection unexpectedly not shows the best performance on all metrics, indicating that some skip connections are not always necessary for detection. Besides, each skip connection (d0, d1, d2, d3, d4) also contributes differently, demonstrating that the independent simple copying



(a) EDRNet (b) CAT-EDNet

Fig. 1. Comparison of the skip connection mechanism between EDRNet and proposed CAT-EDNet.

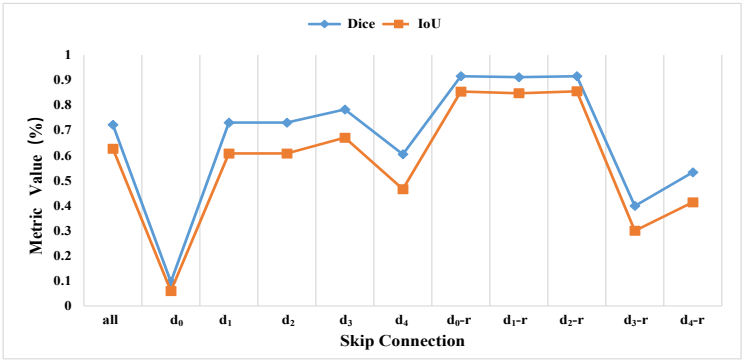


Fig. 2. Analysis of different skip connection layers of EDRNet, "all" is the original EDRNet, "d1" represents only the connection of level one is kept, "d1-r" represents only the connection of level one is removed.

of skip connections has limitations on detection performance. Furthermore, the scores enhance significantly when removing the skip connection d0, d4 (d0-r, d4-r), suggesting that some connections will produce negative influence. We figure out the skip connections issues are essentially derived from the incompatible semantic gap between encoder and decoder rather than the informative encoder features. To address this challenge, the prediction module is a densely supervised encoder-decoder network and a cross-attention transformer (CAT) is embedded to replace the original independent skip connection (Fig. 1 (b)). By collaborating learning of the multi-head structure, the multi-scale features with semantic gap achieve cross-layer communication in channel-wise, and the long-range connection dependency is naturally modeled. Thus, the global-oriented prediction module can completely highlight the salient object and effectively suppress the inconspicuous background.

In addition, we find that the refinement filter RRS_1D (Fig. 5) in EDRNet has slight significance in further optimizing the coarse saliency maps when cooperating with our CAT-based prediction module (TABLE I). The structure of RRS_1D follows a light encoder-decoder style, and the multi-scale encoder features and the up-sampling decoder features are directly concatenated, which cannot effectively utilize the final comprehensive prediction features and inhibits the feature characterization after CAT. Therefore, we equip our residual refinement module with cross-attention (CA) to realize available feature fusion and propose the cross-attention refinement module (CARM) closely after the decoder. By explicitly focusing on temporal features, local-oriented CARM is potential in producing final saliency maps with accurate and smooth boundary.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2022.3165270, IEEE Transactions on Instrumentation and Measurement

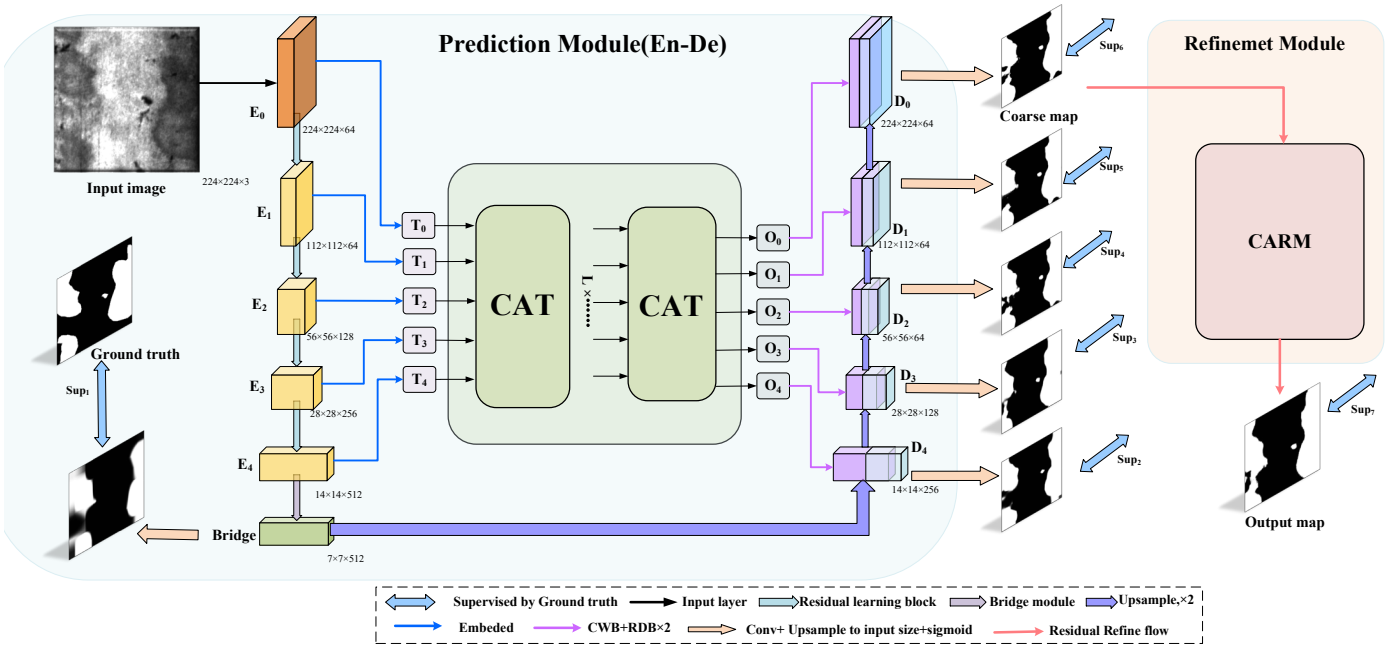


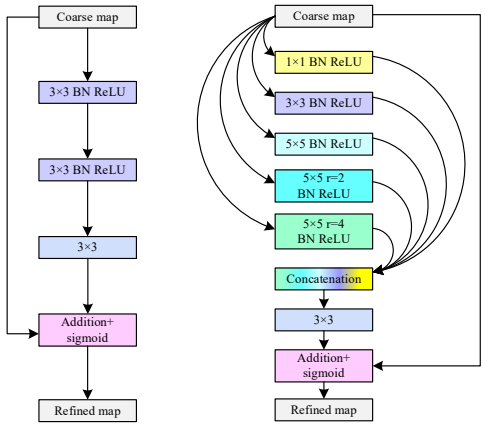
Fig. 3. Architecture of the proposed Cross-Attention Transformer based Encoder-Decoder Network: CAT-EDNet. The original skip connection is replaced by cross-attention transformer (CAT) in the prediction module. While in the refinement network, the CARM denotes the cross-attention refinement module.

B. Vison Transformer

Transformer is an autoregressive language model derived from machine translation，for the strong modelling capabilities and less need for vision-specific inductive bias, it has attracted more and more attention in the field of computer vision [16]. Salient object detection, is essentially segmentation, as a basic but still challenging task also benefits from vison transformer (ViT). *Patch-based Transformer* and *query-based Transformer* are two generally used models [17]. Treating the input image as a patch sequence and feed it into a columnar *Transformer* encoder, *Patch-based Transformer* form different segmentation frameworks with resolution invariance strategy. SETR [18] replaces CNN backbone with transformer encoder and uses multilevel feature aggregation module for pixel segmentation, but it affixes to expensive GPU clusters and extra RAM. TransUNet [19], which can be viewed as a hybrid model of U-NET and transformer, is the first visual transformer for medical image segmentation. To improve transformer performance, Segformer [20] has redesigned a lightweight decoder and embedded a series of measures, such as overlap patch projection. *Query-based Transformer* can aggregate information of each patch more equitably, Panoptic DETR [21] generates a cross-attention module between the object query and encoded features for each object. Through a series of parallel dynamic mask headers with shared queries, QueryInst [22] implements the one-to-one correspondence between mask RoI features and object queries. However, the above transformer architectures are all applied to compensate the strong inductive preference of convolution operations rather than targeting the structure of the segmentation framework, structural redundancy and high computational cost may be involved.

C. Residual Refinement Module

The “coarse map” is determined as the salient map predicted with blurred and noise boundary, uneven regional prediction probability. Thus, the Refinement Modules (RMs) is necessary for the coarse map refining. RMs are usually designed as residual blocks to capture the difference between coarse map and ground truth. Due to the high computational efficiency and less storage, the small 3×3 convolution filters are popular components in RMs. The residual-like block RES, dense-like block DSE, inception-like block INC, residual U-block RSU, are existing typical convolution blocks summarized in [47]. The small receptive field of 3×3 filters in RES and DSE focus on the local details. To extract more global information from shallow high-resolution layers, dilated convolutions are applied in INC. The RSU captures intra-stage multi-scale features with U-structure, has notably smaller computation overhead and improved efficiency.



(a) RRM_LC **(b) RRM_MS**

Fig. 4. Illustration of Residual refinement modules (RRM).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2022.3165270, IEEE Transactions on Instrumentation and Measurement

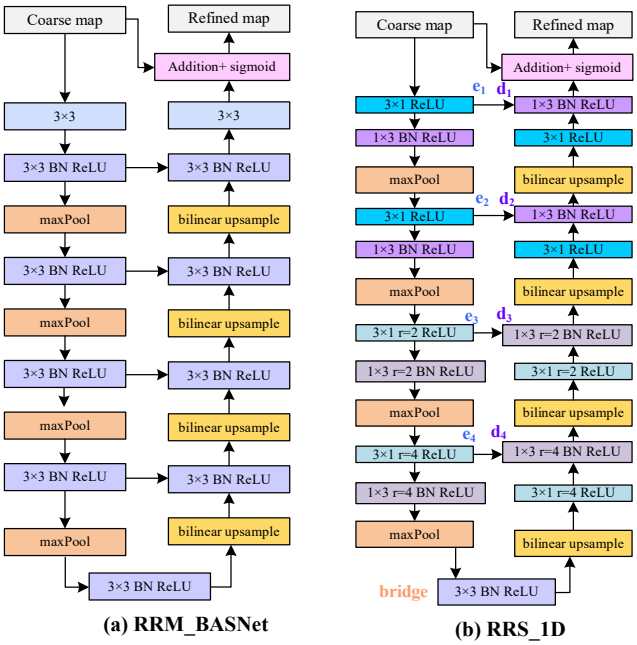


Fig. 5. Illustration of improved residual refinement modules (RRM).

As is shown in Fig. 4, with small receptive field of RES, the residual refinement module based on local context (RRM_LC) [23] is designed for boundary refinement, which is iteratively applied to refine the segmentation probability graph at different scales [24], [25]. Pooling operation will cause details to be lost, so convolutions of INC with different kernel sizes and dilations are configured in multi-scale refinement module (RRM_ MS) [26] to capture multi-scale features. However, these modules are only specialized in capturing shallow information, resulting in less refined maps. To improve the accuracy in refining the regions and boundaries, Qin [27] proposes a novel RSU architecture (see Fig. 5 (a)), It consists of an input layer, an encoder, a bridge, a four-stage decoder and a four-stage output layer. Combining the symmetrical up-sampling and down-sampling operations with skip connections, RRM_BASNet is able to recover more details. Further, to ease computational burden, EDRNet [15] divides the 3 × 3 convolution into two specialized 1D filters (3×1 and 1×3 convolution) and proposes RRS_1D (see Fig. 5 (b)), dilated convolutions (r=2, r=4) are also employed in it to obtain a larger receptive field.

III. Cross-Attention Transformer Based Encoder-Decoder Network

As illustrated in Fig. 3, our CAT-EDNet belongs to a predict-refine framework, the prediction module, embedded with a cross-attention transformer (CAT), is a densely supervised encoder-decoder network. First, the coarse probability maps are learned from input images through the prediction module. And then, the output map is finally generated by learning the residuals between the coarse map and the ground truth through cross attention refinement module (CARM). In addition, the whole process is guided by the hybrid loss to learn three levels (pixel-, patch-, map-) features.

A. Prediction module

1) Encoder

Large scale features obtained from deep low resolution feature maps, can provide more semantic information while sacrificing the spatial resolution. Due to the skip connections and stepwise up-sampling are effective in recovering high-resolution probability map, the encoder-decoder like architectures achieve significant performance in segmenting edge or slender structures. As is illustrated in Fig. 3, for the encoder, we introduce an input layer, four residual learning blocks, a bridge module. The input image is first fed into the input layer, which has 64 3 ×3 convolution filters with stride of 2, and the output map E_0 has the same spatial resolution with the input image, the adaptability enables the network to obtain higher resolution feature maps at earlier levels. To enlarge the receptive fields, the four residual learning blocks, which inherit from ResNet-34 (conv2-3, conv3-4, conv4-6, conv5-3), are improved by previously adding a max pooling operation with kernel size 3 ×3 and stride 2. And the resolution of E_1, E_2, E_3, E_4 are decreasing step by step when successively passing the blocks of 64, 128, 256, 512 layers. To accurately locate the object region and completely segment the defect, a bridge module is laid between encoder and decoder to capture richer global semantic information. It consists of three 512-channel convolution layers with dilated (dilation rate= 1, 2, 4) 3 ×3 filters, and the first convolution use stride 2 to maintain the same resolution with the original ResNet-34. Noted that during the whole encoder process, after each convolution output, the batch normalization layer is cooperated with a ReLU activation function to alleviate gradient disappearing and enhance the nonlinear characterization ability of the model.

2) Cross-attention transformer

The cross-attention transformer (CAT), which has strong long range dependency modeling capability, is applied to fuse the multi-scale encoder features of skip connection layers. Inspired by [28], ***Multi-scale feature embedding***, ***Multi-head cross attention*** and ***Multi-layer perception*** (MLP) are "3M_s" equipped in the CAT.

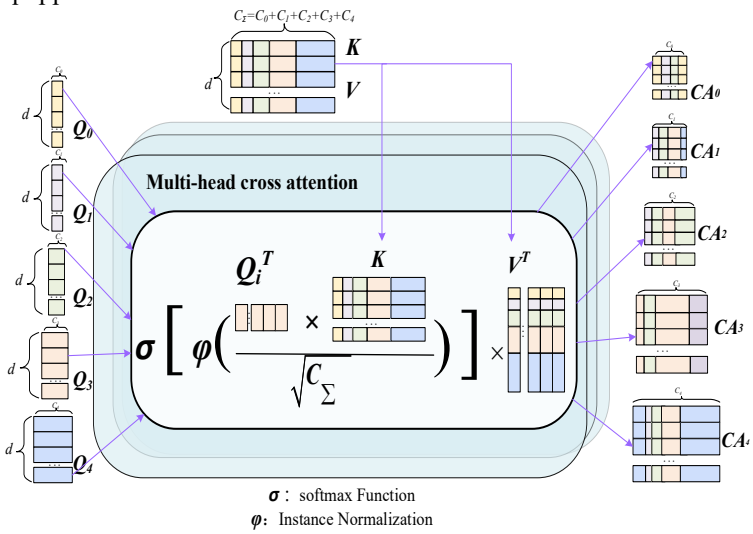


Fig. 6. The multi-head cross attention mechanism.

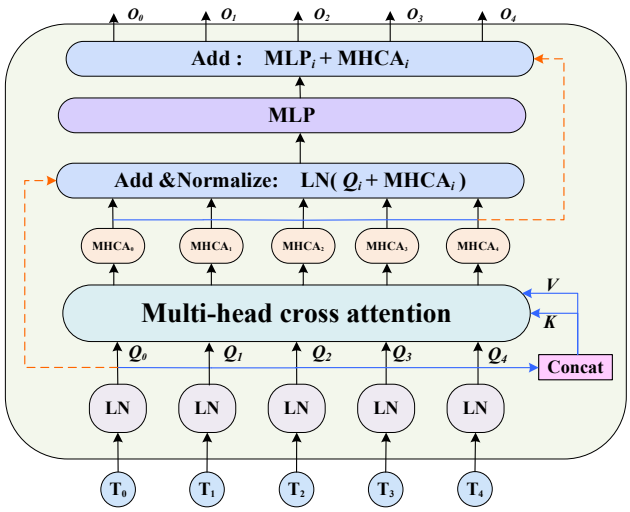


Fig. 7. The structure of cross-attention transformer.

As is shown in Fig. 3, the input image $I \in \mathbb{R}^{H\times W\times C}$ (H, W, C is the height, width, channel number, respectively) is imported to extract the multi-scale five-level feature $E_i \in \mathbb{R}^{H/2^i\times W/2^i\times C_i}$ (i=0, 1, 2, 3, 4), the channel dimensions are C_0=64, C_1=64, C_2=128, C_3=256, C_4=512, respectively. In the ***multi-scale feature embedding*** process, to map the same area feature representation of the five-scale encoders, we reshape $\boldsymbol{E_i}$ into sequential 2D patches with size $\boldsymbol{P/2^i}$ (i=0, 1, 2, 3, 4; $\boldsymbol{P}$=32), and naturally form the different token $\boldsymbol{T_i}$. Finally, the key and value are obtained by concatenating the five layers $\boldsymbol{T_i}$ (i=0, 1, 2, 3, 4), represented as $T_\Sigma = Concat(T_0, T_1, T_2, T_3, T_4)$.

We can see from Fig. 7, without adding extra computation, the ***multi-head cross attention*** module is introduced to aggregate the relationships and dependencies of multi-scale encoder embedding features. And then a simple position-based ***MLP*** is followed to refine feature representation. The residual structure is to guarantee the scalability of the model. The queries in the Fig. 6**.** are learned by:

$$Q_i = T_i W_{Q_i}, K = T_\Sigma W_K, V = T_\Sigma W_V \tag{1}$$

Where the quires $Q_i \in \mathbb{R}^{C_i\times d}$, key $K \in \mathbb{R}^{C_\Sigma\times d}$ and value $V \in \mathbb{R}^{C_\Sigma\times d}$ are transformed by the weight parameters $W_{Q_i} \in \mathbb{R}^{C_i\times d}$, $W_K \in \mathbb{R}^{C_\Sigma\times d}$, $W_V \in \mathbb{R}^{C_\Sigma\times d}$, respectively.

To measure the similarity, the cross-attention value $\boldsymbol{CA_i}$ in the Fig. 6 is calculated by:

$$CA_i = \sigma\left(\varphi\left(\frac{Q_i^T K}{\sqrt{C_\Sigma}}\right)\right)V^T = \sigma\left(\varphi\left(\frac{W_{Q_i}^T T_i^T T_\Sigma W_K}{\sqrt{C_\Sigma}}\right)\right)W_V^T T_\Sigma^T \tag{2}$$

Where $\boldsymbol{Q_i^T K}$ represents the correlation score of channel-based similarity maps rather than patch-based, normalized by dividing $\sqrt{C_\Sigma}$ to making the gradient more stable during training. The $\boldsymbol{\sigma}$ **(.)** denotes softmax function, which converts the score vector to probability value. And $\boldsymbol{\varphi}$ **(.)** is the instance normalization operation to propagate the gradient more smoothly. For the multi-head cross attention module of N head, the output $\boldsymbol{MHCA_i}$ is expressed as:

$$MHCA_i = \frac{CA_i^1 + CA_i^2 + ,..., + CA_i^N}{N} \tag{3}$$

To prevent training degradation and accelerate model training speed, ***Add & Layer Normalize*** operation is followed. Finally, the output $\boldsymbol{O_i}$ is obtained by performing ***MLP*** and residual operation:

$$O_i = MHCA_i + MLP(LN(Q_i + MHCA_i)) \tag{4}$$

In addition, the L-layer CAT in Fig. 3 is designed by repeating the operation in (4) $\boldsymbol{L}$ times. In this paper, $\boldsymbol{N}$ and $\boldsymbol{L}$ are both set to 5. By up-sampling operation combined with a convolution layer, $O_i \in \mathbb{R}^{H/2^i\times W/2^i\times C_i}$ is finally reconstructed to integrate with decoder features.

3) Decoder

To eliminate the semantic ambiguity between CAT and decoder, the features $\boldsymbol{O_i}$ processed by CAT are fused with decoder features by channels weighted block (CWB) and residual decoder block (RDB) in [15], which can guide the channels to gradually recover the saliency information.

1) *CWB* is depicted in Fig. 8 (a), the current transformer features $\boldsymbol{O_i}$ containing long range dependencies, are concatenated with the next decoder feature $\boldsymbol{Y}$ **(**In Fig. 8 (c), noted that when i=4, $\boldsymbol{Y}$= *Bridge*; when i<4, to keep the resolution the same, $\boldsymbol{Y}$ is obtained by up-sampling $\boldsymbol{D_i}$). And then, the global average pooling (GAP) operation is applied to learn global context information, which is beneficial to predict salient defect region and suppress the background noise. Besides, the followed two 1×1 convolution layers can reduce dimension, keeping the model lightweight. **PReLU** is embedded to enhance generalization ability while the weight vector $\boldsymbol{W}\in$ **[0,1]** is obtained by sigmoid function. Finally, the residual-learned output **Z** is generated by element-wise summing the weighted $\boldsymbol{O_i}$ and initial $\boldsymbol{Y}$.

2) *RDB* in Fig. 8 (b) is used to gradually recover the encoded multilevel information. After CWB, **Z** is first fed into parameter-fewer channel shuffle unit to achieve higher detection efficiency and promote optimization. And then, pass 3×3 convolution layer with batch normalization **BN** and **PReLU**, 1×1 convolution is closely followed to limit model complexity and interact the cross-channel information.

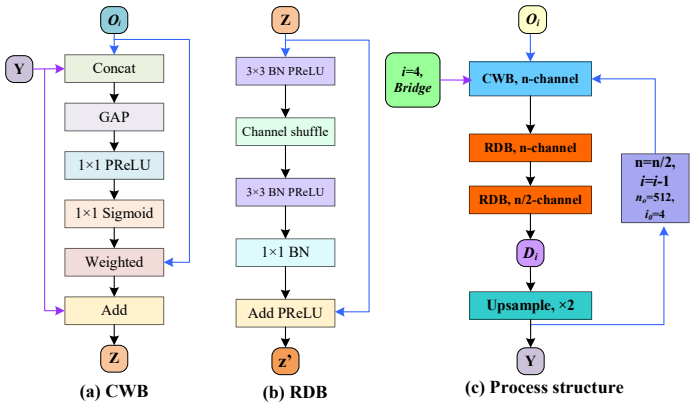


Fig. 8. Detailed structure of the components in decoder.

The detail of the decoder in Fig. 3 is expanded in Fig. 8 (c), after a ***CWB*** and two ***RDB*** operations, the five-level decoder features D_i are formed in each stage, in addition, producing five side-output saliency maps deeply supervised by ground truth, which can guide the network to learn correct defect region.

B. Refinement module

The deep supervision mechanism can be reflected in Fig. 3, where five supervision signals are imposed in the bridge and five-stage decoder output saliency maps. The last coarse map contains the most comprehensive semantic information, obtaining the highest detection accuracy. However, there is a lack of refinement of boundary and region details. To further optimize the detection effect, we propose the feature-wise *Cross-Attention Refinement Module* (CARM). The aim is to learn the residuals S_{res} between S_{coarse} and ground truth.

$$S_{refined} = S_{coarse} + S_{res} \tag{5}$$

Inspired by II.C, our CARM continues to be a lighter encoder-decoder structure, which is inherited by RRS_1D. When cooperating with our CAT-based prediction module, we find the RRS_1D bring slight performance improvement, which is due to the simple concatenating operation between encoder and decoder, cannot effectively fuse features with inconsistent semantics. Therefore, we introduce the cross-attention (CA) to better extract the feature with fine characterization ability. As is shown in Fig. 5 (b), the four-stage U-block structure is the repetition of each stage which consists of two specialized 1D filters (64-channel 1×3, 3×1 convolution) and max pooling or bilinear up-sampling operation. The 1D filters are computationally efficient, and max pooling is for down-sampling, making the network deeper while reducing the computation. Up-sampling is used to match the feature dimension. Besides, the larger receptive field is obtained by dilated convolutions (r=2, r=4) and the bridge unit is composed by 3×3 convolution layer of 64 channels. From Fig. 9, taking the ***i-th*** stage encoder output $e_i \in \mathbb{R}^{H_i\times W_i\times C}$ and decoder output $d_i \in \mathbb{R}^{H_i\times W_i\times C}$ (i=1,2,3,4; $H_i = H/2^{i-1}$, $W_i = W/2^{i-1}$; C=64) in Fig. 5 (b) as the input of CA, the global average pooling (GAP) is performed to achieve spital squeeze, and the ***K-th*** channel will transform into a globally distributed value $G(X)=\frac{1}{H_i\times W_i}\sum_{i=1}^{H_i}\sum_{j=1}^{W_i}X^k(i,j),\in\mathbb{R}^{C\times1\times1}$. Then, to model dependencies between channels, the attention mask is generated by:

$$M_i = L_1 \bullet G(e_i) + L_2 \bullet G(d_i) \tag{6}$$

Where $L_1 \in \mathbb{R}^{C\times C}$ and $L_2 \in \mathbb{R}^{C\times C}$ denote the weights of single linear layers and sigmoid function. The original feature recalibration is completed by assigning channel importance to each pixel of each channel and formed e_i'. Finally, the fused feature is obtained by concatenating the masked feature e_i' and encoder feature d_i. After refinement module, the output saliency map is the finally results of our CAT-EDNet, which is also supervised by ground truth.

C. Hybrid loss

The difficulty of the salient defect detection of strip steel lies not in the obvious salient object, which has high contrast with the background, but in the camouflaged defect objects, which has similar appearance with the background. Besides, capturing complex structure with complicated boundary is also challenging. To make the network perceive the hard object, the hybrid loss is used during the training process by guiding the network learn pixel-, patch-, map- level hierarchy representation. Compared with the current methods focused more on high regional accuracy, the hybrid loss has more robust and competitive performance in high spital accuracy of boundary and fine structures.

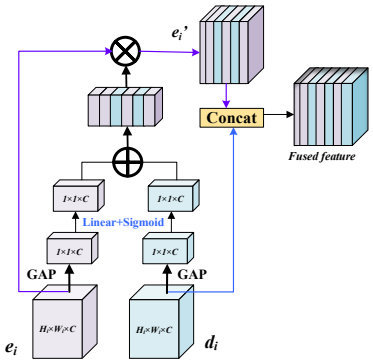


Fig. 9. Detailed structure of the CA.

The training loss is defined as the weighted sum of all the losses supervised by ground truth, including bridge loss, five side-output loss, refinement loss:

$$L_{total} = \frac{1}{B}\sum_{k=1}^{7}\alpha_k l^{(k)} \tag{7}$$

Where α_k is the weight of k-th loss，B denotes the batch size. In addition, the hybrid loss $l^{(k)}$ is formulated as:

$$l^{(k)} = l_{bce}^{(k)} + l_{iou}^{(k)} + l_{ssim}^{(k)} \tag{8}$$

Where l_{bce}， l_{iou} and l_{ssim} represent binary cross entropy (BCE) [29], intersection-over-union (IoU) [30], structural similarity (SSIM) [31], respectively.

The BCE loss is measured in pixel-level. It does not consider neighborhood labels, and gives equal weight to foreground and background pixels. This facilitates convergence at all pixels and ensures a relatively good local optimization, also maintains smooth gradients for all pixels. Which can be defined as:

$$l_{bce} = -\sum_{(r,c)}[G(r,c)\log(s(r,c)) + (1-G(r,c)\log(1-(s(r,c)))] \tag{9}$$

Where $G(r,c)$ is the binary ground truth label of pixel (r,c), 0 is the background while 1 denotes defect object. $S(r,c)$ represents the predicted probability of corresponding pixel.

However, l_{bce} usually results in fine structure but blurred boundaries of foreground, therefore, to pay more attention to boundary and foreground region, by considering the local neighborhood of each pixel, the patch-level SSIM originally designed to capture structural information in an image is introduced. It gives higher weight to pixels in the buffer area between foreground and background, such as boundary and fine structure. For two corresponding patches x and y of size $N\times N$ cropped from predicted saline map and ground truth，$\boldsymbol{X}=\{\boldsymbol{x}_j : j=1,\ldots,N^2\}$, $\mathbf{Y}=\{\boldsymbol{y}_j : j=1,\ldots,N^2\}$,the SSIM is calculated by:

$$l_{ssim} = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{10}$$

Where μ_x, μ_y and σ_x, σ_y respectively represent mean value and variance of x and y. $C_1=0.01^2$ and $C_2=0.03^2$ are empirically set to avoid nan.

Larger regions contribute more to map-level IoU, so models trained by IoU can predict relatively homogenous and more confident probabilities for the larger prospective regions. It is formulated as:

$$l_{iou}=1-\frac{\sum_{r=1}^{H}\sum_{c=1}^{W}s(r,c)G(r,c)}{\sum_{r=1}^{H}\sum_{c=1}^{W}[s(r,c)+G(r,c)-s(r,c)G(r,c)]} \tag{11}$$

Where $G(r,c)$ and $S(r,c)$ mean the same as in l_{bce}. However, the model often involves false negatives on the fine structure due to the biased preference for foreground regions.

By implicitly injecting fine structure optimization goal during training process, the three-level losses are fused to formulate the hybrid loss. Thus, the pixels, foreground defect objects and boundaries are comprehensively considered.

IV. Experiments

A. Experimental Setup

1) Implementation Details

The experiments are all performed on 12GB Nvidia Titan XP GPU, 2.2GHz Intel Xeon E5-2630 CPU and 64GB RAM. From SD-saliency-900 [15], we randomly selected 540 (180×3) images of the three defects: inclusion (In), patch (Pa), scratch (Sc). And to simulate noise interference, the collected 270 (90×3) images from the previous 540 images are randomly added different levels of Gaussian and salt & pepper noise. So the training dataset containing 810 images are constructed, some samples can be visualized in Fig. 10. Noted that to weaken data noise and strengthen model stability, each training sample (200×200) is first resized to 256×256, randomly cropped to 224×224, and then normalized by dividing by the standard deviation 0.2437 after subtracting mean value 0.4669. The parameters of our encoder network are initialized by employing ***He*** strategy [32]. Besides, to obtain a fast convergence speed, RMSprop [33] optimizer is applied during the training process, where the initial learning rate is set to 0.001 and alpha is set to 0.9. We also configure the CAT hyperparameters as follows: the embedding dropout rate=0.1, attention dropout rate=0.1, channel dimension ratio=4, KV size=1024, patch sizes=[32,16,8,4,2], heads number=5, layers number=5. Taking about 8 hours, with the batch size of 8, our model converge after 70-K iterations. In addition, the test samples are also randomly added with varying levels of noise, only resized to 256×256, and then fed into the trained network. Using bilinear interpolation, the output saliency maps are finally resized back as the original input image size.

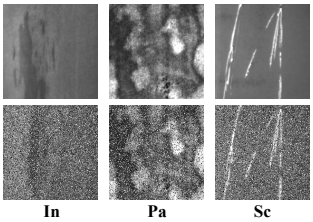


Fig. 10. Some samples of our training dataset. The corresponding noisy images are shown in the second row.

2) Evaluation Metric

We adopt six metrics to evaluate the salient detection performance of our model. **(1) Structure Measure (SM)** [34] contains the region-aware and object-aware structural S_r and S_o, the overall structural information of object is captured by S_r while S_o compares the global distribution of the foreground and background. **(2) Weighted F-measure (WF)** [35] is the weighted harmonic average of the precision and recall, comprehensively evaluating the influence of dependency, interpolation and equal-importance. **(3) Mean absolute error (Mae) [**36] measures the average difference of pixels between predicted salient map and ground truth. **(4) Enhanced alignment measure** (**Eam**) [37] jointly captures the image-level statistics and local pixel matching information. **(5) Dice coefficient (Dice)** [38] is an original measure of set similarity, commonly used to calculate the similarity of probability maps and ground truth in the medical segmentation field. **(6) IoU (Intersection over Union)** [39] globally evaluates the images based on class calculation.

B. Ablation Studies

In this section, to validate the effectiveness of the core components configured in our model, three groups of ablation experiments are performed: architecture analysis, loss ablation and the research of number of queries and keys.

1) Architecture Analysis

To demonstrate the effectiveness of our proposed CAT-EDNet, we conduct a series of experiments under the hybrid loss l_{total} to quantitatively compare our model with the related components. The EDRNet without CBAM is taken as our baseline, the original independent skip connection is replaced by adding CAT. As illustrated in TABLE I, CAT can bring significant performance improvement, which is beneficial from its strong long-range dependencies modelling ability. Then we progressively test the effect of CAT on other modules by removing the corresponding decoder modules CWB, RDB and refinement module RRS1D, respectively. We find that removing CWB severely degrades performance while the scores are slightly enhanced when subtracting RDB and RRS1D. However, the qualitative results in Fig. 12 show that removing RDB will lose the detailed boundary information, indicating its ability in gradually recovering encoded multilevel features. Therefore, RDB is retained and RRS1D is replaced by CARM to effectively fuse features with inconsistent semantics when cooperated with CAT, which is our CAT-EDNet. The

TABLE I

Ablation study of different architectures.

Architecture	Metric(%)					
	SM↑	WF↑	Mae↓	Eam↑	Dice↑	IoU↑
EDRNet [15]	78.34	78.04	2.71	85.96	72.24	62.69
EDRNet (CAT_+)	90.85	90.32	1.43	96.96	91.66	85.36
EDRNet (CAT_+CWB_-)	50.99	36.11	10.0	48.17	25.12	15.80
EDRNet (CAT_+RDB_-)	90.94	90.45	1.42	97.14	91.80	85.74
EDRNet (CAT_+RRS1D_-)	90.87	89.94	1.43	97.19	91.75	85.69
CAT-EDNet	**93.51**	**93.63**	**1.15**	**97.95**	**94.27**	**90.31**
CAT-EDNet ($CBAM_+$)	90.76	89.97	1.47	96.98	91.60	85.46

The subscript "+" indicates the network architecture configures this module, while "-" represents using CAT-Net as training model but removed the module. Noted that the EDRNet is all trained without CBAM module.

TABLE II
Ablation study of different loss.

Loss	Metric(%)					
	SM↑	WF↑	Mae↓	Eam↑	Dice↑	IoU↑
l_{bce}	90.86	90.38	1.44	97.22	91.40	85.22
l_{iou}	90.83	91.16	1.52	97.79	91.04	84.65
l_{ssim}	90.03	89.68	1.53	97.17	90.67	84.19
$l_{bce}+l_{iou}$	90.55	90.78	1.46	97.61	91.46	85.33
$l_{bce}+l_{ssim}$	90.51	90.65	1.46	97.32	91.04	84.74
L_{total}	**93.51**	**93.63**	**1.15**	**97.95**	**94.27**	**90.31**

TABLE III
Ablation study of the number of queries and keys.

Queries /Keys	Metric(%)					
	SM↑	WF↑	Mae↓	Eam↑	Dice↑	IoU↑
Q0	91.28	90.64	1.38	97.21	91.90	86.00
Q1	90.89	89.03	1.46	96.85	91.35	85.33
Q2	90.96	89.39	1.44	96.82	91.53	85.44
Q3	91.09	90.07	1.41	96.79	91.72	85.65
Q4	91.14	90.35	1.14	97.50	91.58	85.49
Q01	91.23	89.58	1.44	97.03	91.69	85.69
Q23	91.39	90.38	1.38	97.20	91.86	85.98
Q012	91.11	90.17	1.41	97.15	91.67	85.68
Q123	91.09	89.90	1.45	97.00	91.65	85.58
Q1234	91.29	90.80	1.36	97.40	91.76	85.81
Ours	**93.51**	**93.63**	**1.15**	**97.95**	**94.27**	**90.31**
K0	91.17	89.69	1.42	97.26	91.57	85.57
K01	91.19	89.79	1.42	97.09	91.63	85.70
K012	90.74	89.24	1.48	96.85	91.39	85.14
K0123	91.17	89.84	1.41	97.36	91.62	85.62
K23	91.21	89.98	1.41	97.34	91.82	85.84
K123	91.06	90.29	1.44	96.97	91.50	85.50
K1234	91.21	90.13	1.40	97.28	91.76	85.77

metric values and visual effect both reveal our CAT-EDNet can further optimize the salient detection results. In addition, we validate the inefficiency of the CBAM module embedded in the encoder feature extraction process of our approach, the CBAM introduces excessive attention and makes the self-defeating visual effect.

2) Loss Analysis

A set of experiments over different losses are conducted based on our CAT-EDNet. The results in TABLE II indicate that the hybrid loss l_{total} achieves the most excellent performance by guiding the network learn pixel-, patch-, map-level hierarchy representation. Compared to the commonly used single l_{bce}, the SM, WF, Dice and IoU are increased by 2.23%, 3.25%, 2.87%, 5.03%, respectively. We visually compare the impact of different loss on salient detection of defects (In, Pa, Sc) in Fig. 13. Suppressing errors by giving a prediction value of around 0.5 near the boundary, the l_{bce} generates the foreground with blurred boundary. l_{iou} places more emphasis on larger foreground region, producing false negative in relatively finer structure. l_{ssim} ignores the accuracy, which is manifested in could not clearly separate different parts close to each other and characterizing boundary details too smoothly. In addition, combined the l_{bce} with l_{iou} or l_{ssim} still cannot effectively improve the salient detection quality. By contrast, the hybrid loss can highlight the complete object and optimize the boundary localization, simultaneously.

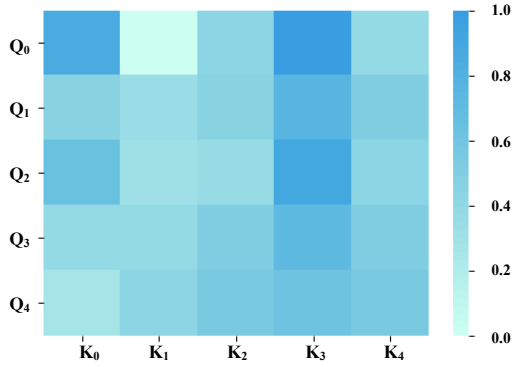


Fig. 11. Similarity matrix of cross attention distributions.

3) Number of Queries and Keys Analysis

The CAT module in the architecture ablation section shows its effectiveness in greatly enhancing the defect integrity. The extracted multi-scale encoder features achieve cross-layer communication in CAT through its multi-head structure. Thus, the number of quires is set to 5 and the keys are obtained by concatenating the five-stage representation. As shown in TABLE III, a series of experiments are conducted with different amount of skip connections between encoder and decoder. When compared with various queries representing different encoder levels, the key vector is fixed as five-scale features. We find Q0, which focuses on the spatial boundary details of object reconstruction, is more confidently associated with the salient detection. and the performance behaves consistent improvement with the increase of our learned encoder levels Q01234 by allocating bigger weights to the shallow low-layer. In addition, by keeping the queries fixed and varying the keys, as visualized in the Fig. 11, we observe that K3 has more confident correlation, which is consistent with the skip connection "d_3" in Fig. 2. Besides, by introducing more channel information, the performance is enhanced with our concatenated all multi-scale features. Fig. 14 provides qualitative comparison between different number of quires and keys, which also indicate transforming more scales of features to queries is helpful to accurately represent object and finely capture boundary details.

C. Comparison with the State-of-the-art

To demonstrate the overall salient detection performance of our proposed CAT-EDNet, we compare it with *twelve* state-of-art models, including BASNet [27], PiCANet [40], UCtransNet [28], PoolNet [41], CPD [42], EGNet [43], SINet [44], PFANet [45], EDRNet [15], RSNet [46], U^2-Net [13] and image matting [14]. To make a fair comparison, we use the originally released codes and published setting, and retrained all the models on the same training dataset as ours.

1) Visual Comparison

We visualize the salient detection results of the comparable models in the Fig. 15, our CAT-EDNet can generate better salient maps in different challening cases. For the small defects with relatively low contrast, which will be easily interferneced by background clutter (1st, 2nd, 3rd row), some models (EGNet, SINet, PFANet) will either produce confusing false positives in the background region and have ambigious perception to the

object, or ignore the the tiny defects easily swallowed by noise (EDRNet, RSNet). However, embedded with global-oriented CAT, our model can accurately distinguish the whole defect object without losing any detailed part. Besides, our approach also has competitive performance in large patch defects with complex background and complicated boundary (4th, 5th, 6th row), PiCANet predicts the object as scattered mass, EGNet, SINet, PFANet identify the near parts that do not intersect as adhesions, CPD, EGNet also output low contrast salient maps with haloes-like boundary effect. By contrast, our CAT-EDNet has potential in uniformly highlighting the complete defect with coherent distinct boundary. It is remarkable that for the slender defects with very finer boundary (7th, 8th, 9th row), only BASNet, UCtransNet, U^2-Net can refine the boundary distribution, however, configured with local-oriented CARM, our model pays more attention to the finer boundary representation while preserving detailed shape information. In addition, we further compare the local details captured by the several superior models in Fig. 16. As we can observe, other models either prone to produce over-smooth boundary, missing the zig-zag wrinkle like GT, or insufficiently segment the fragile structure, predicting low-resolution saliency results. But our model shows extra promising capability in extracting high--resolution local details. Contrary to our expectation, when feeding the trained image matting model uniform background to simulate the unknown actual production line, the background texture becomes complicated due to the fluctuating noise and lighting, resulting in poor quality matting (see Fig. 15 (m)).

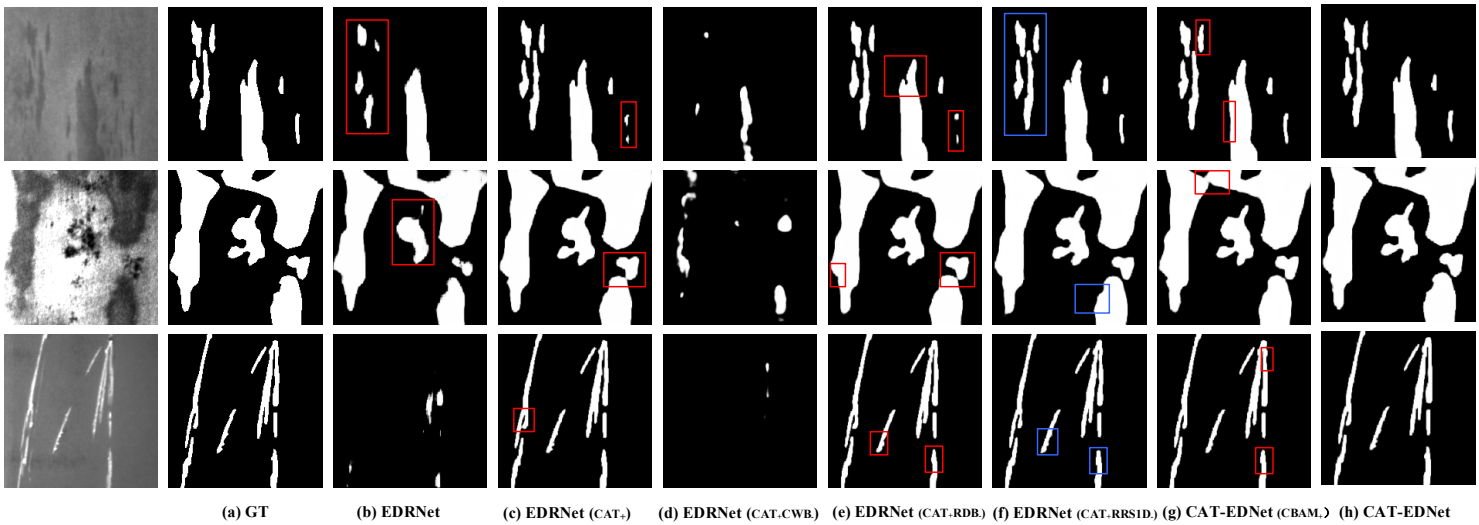


Fig. 12. Qualitative comparison of different configures in the ablation study. The EDRNet is without CBAM and all results are under l_{total}.

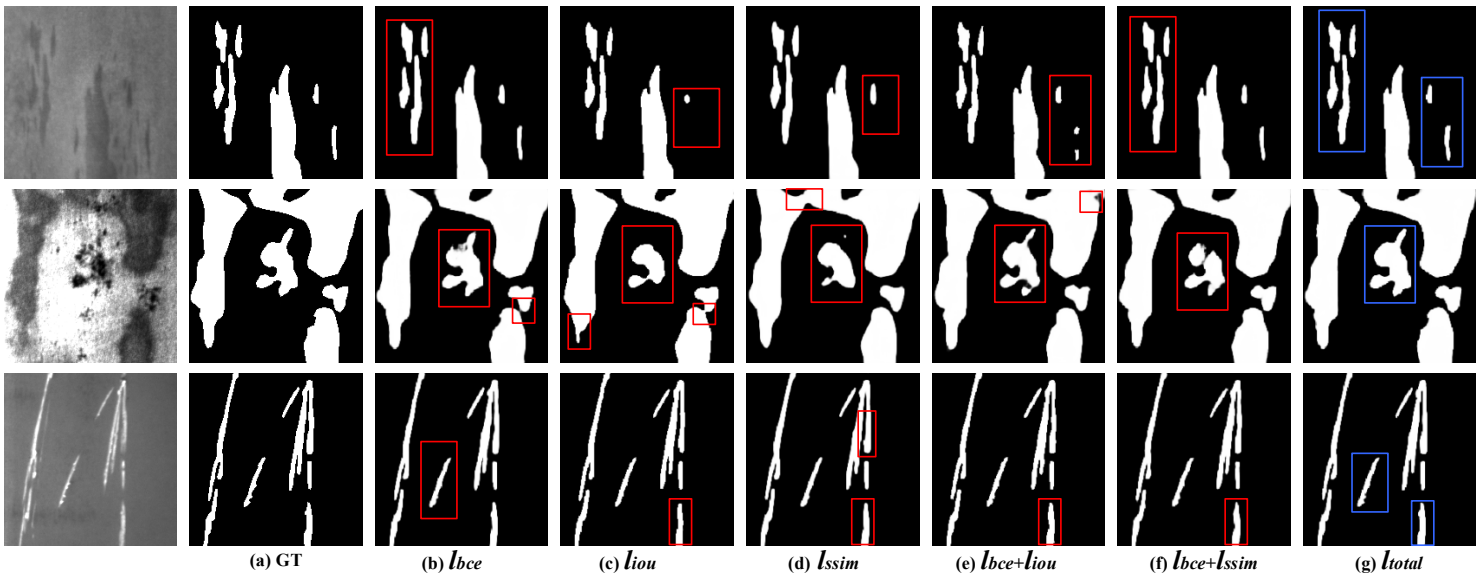


Fig. 13. Qualitative comparison of our CAT-EDNet under different losses in the ablation study.

TABLE IV

Comparison of running time and model size of different methods.

Methods	BASNet	PiCANet	UCtransNet	PoolNet	CPD	EGNet	SINet	PFANet	EDRNet	RSNet	Matting	U^2-Net	ours
Time(fps)	27.92	23.01	22.29	30.91	30.50	27.12	28.87	33.90	26.22	35.64	**40.41**	34.30	28.54
Size(MB)	348.6	189	797.6	278.6	192.2	447.1	196.7	131.1	157.6	**99.1**	322.7	176.4	312.7

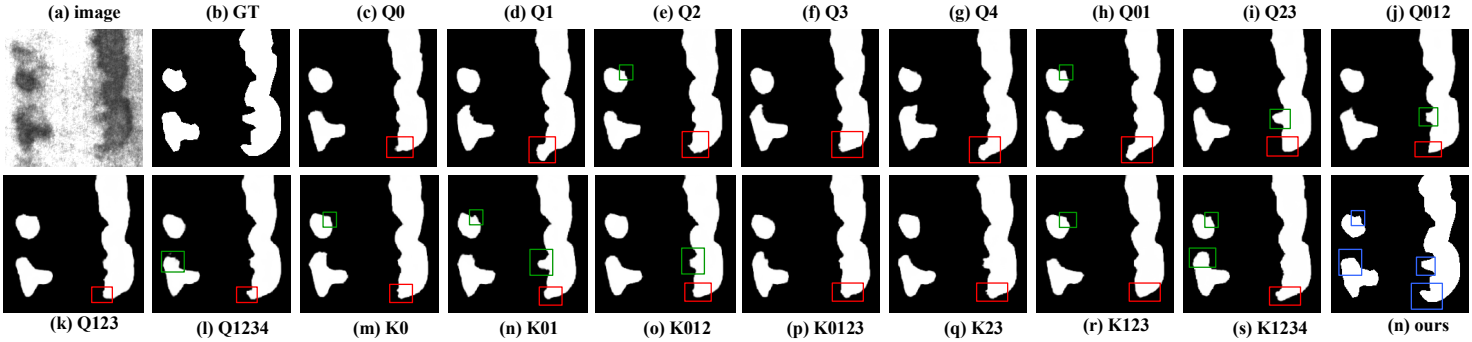


Fig. 14. Qualitative comparison of different number of queries and keys.

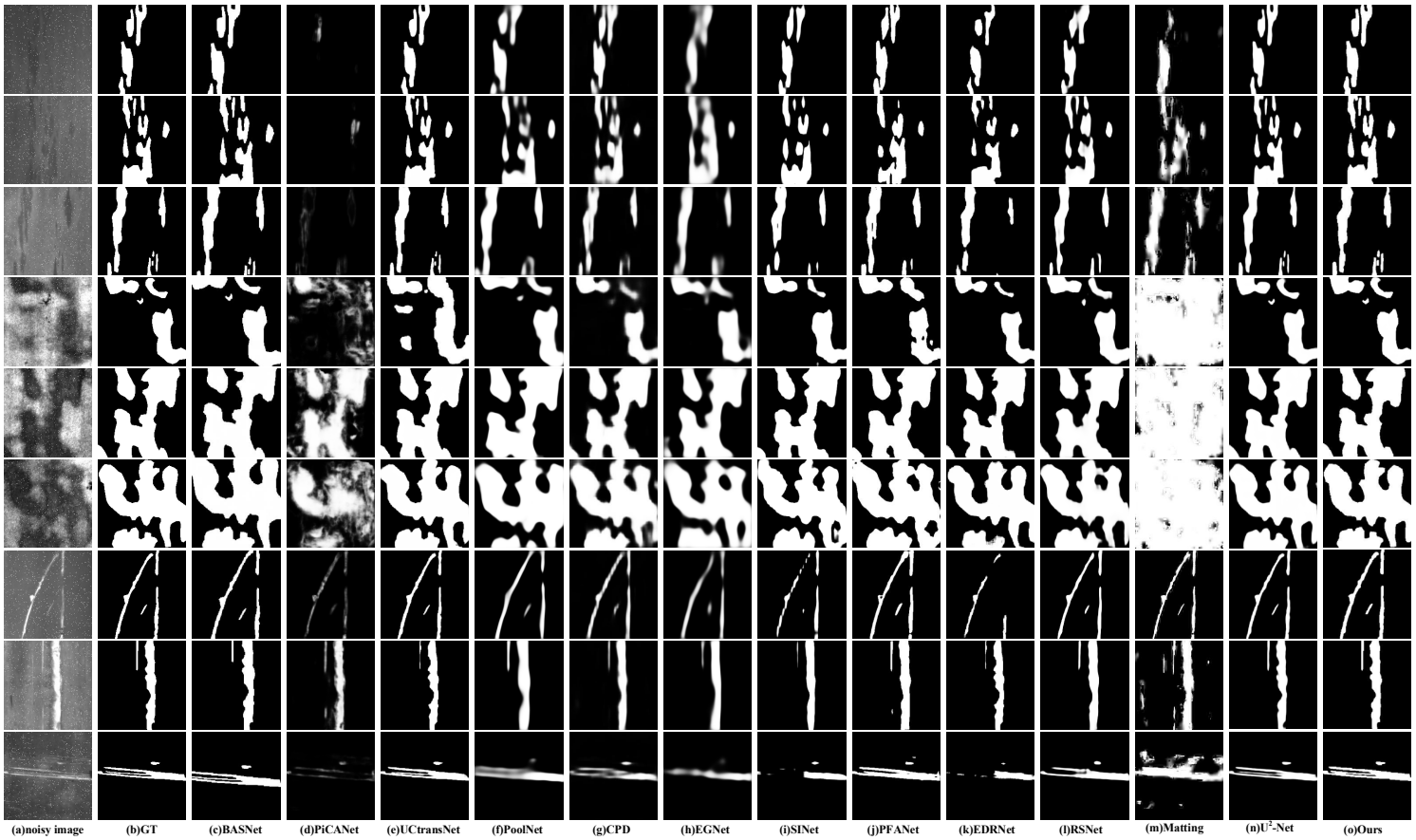


Fig. 15. Visual comparison of saliency maps. The noisy images are obtained by adding random noise.

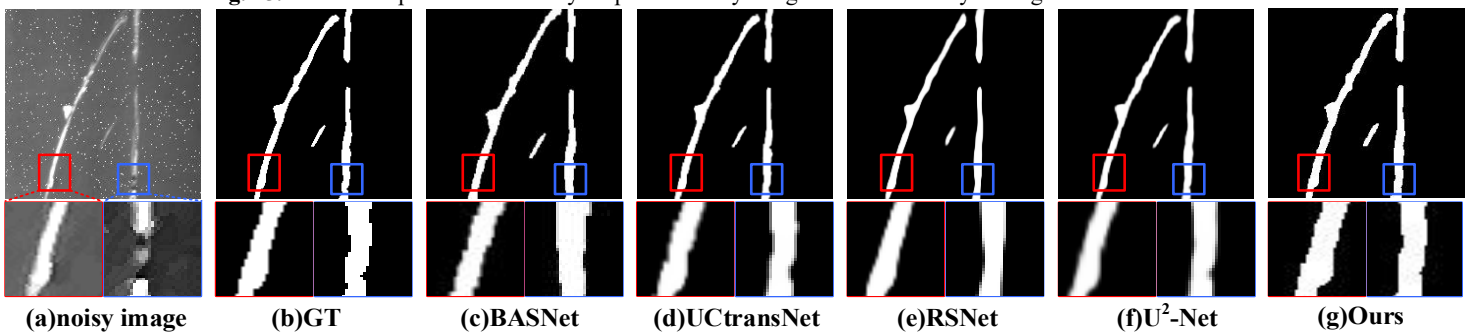


Fig. 16. The visualization comparison of local details predicted by previous state-of-the-arts.

2) Quantitative Comparison

The quantitative results are reported in TABLE V, our CAT-EDNet achieves consistent improvements in terms of nearly all metrics except Mae. The BASNet and EDRNet are both deeply supervised two-stage predict-refine framework, performing superior to other methods, which proves the strong boundary-aware ability of the encoder-decoder network and residual refinement module. However, due to the complicated background texture interfered by noise and illumination, the two-stage image-matting technique is no longer suitable for the salient defect detection of strip steel. For other one-stage methods, various multi-scale feature aggregation strategies are

introduced. U^2-Net yields higher quality by running two-level nested U-structure. RSNet has satisfactory results by employing reverse attention block to guide learning residual in each side-output. UCtransnet also obtains competitive results by combining the channel transformer module into U-Net. Our approach improves the predict-refine architecture by embedding CAT in multi-scale spatial domain to guarantee defect integrity, by introducing CARM in temporal-domain to further optimize defect boundary details, thus, the improvements of our CAT-EDNet against the above six models are significant. Noted that the all models suffer from the frequent noise occurred testing environment. PiCANet generates contextual attention map for each pixel with the only prominent index Eam of 83.80%. PoolNet which has two pooling-based modules to progressively refine high-level semantic features, has weak noise-resistant ability of 11.41% WF. CPD framework focuses on fast salient detection by discarding larger resolution features of shallow layers, also obtaining poor metrics. The edge guidance network EGNet fails to refine the coarse noisy boundary with 10.09% WF. SINet is specially designed to identify objects having high intrinsic similarities with their surroundings, not applicable for camouflaged objection detection with background clutter. PFANet is also sensitive to noise with 10.73% WF when extracting high context-aware pyramid feature. By contrast, as the metrics reflected, our CAT-EDNet can filter out irrelevant background noise, which is first roughly screened by global CAT, and then further fine-filter is achieved by local CARM.

TABLE V

Comparisons with *twelve* state-of-the-arts in terms of *six* quantitative metrics.

Methods	Metric(%)					
	SM↑	WF↑	Mae↓	Eam↑	Dice↑	IoU↑
BASNet	93.21	92.81	**1.04**	97.69	94.20	90.25
PiCANet	67.02	47.06	8.54	83.80	36.12	29.30
UCtransnet	93.09	92.36	1.24	98.06	93.59	89.27
PoolNet	41.56	11.41	19.76	62.69	13.48	8.71
CPD	41.89	11.08	19.70	63.40	13.24	8.62
EGNet	40.81	10.09	20.12	61.26	12.06	7.50
SINet	42.25	8.00	27.4	61.91	11.52	7.21
PFANet	40.56	10.73	20.4	60.42	12.50	7.80
EDRNet	77.79	78.05	3.08	85.19	72.01	62.39
RSNet	89.81	87.73	1.85	96.07	89.21	81.69
Matting	40.81	10.71	19.77	61.26	12.21	7.86
U^2-Net	90.56	88.00	1.33	96.97	90.68	85.87
Ours	**93.51**	**93.63**	1.15	**97.95**	**94.27**	**90.31**

3) Time Efficiency

The interference time and model size are summarized in TABLE IV. Our model takes 28.54 fps interference time with size 312.7 MB. Compared to the RSNet, image matting and U^2-Net, which are specially focused on model weights and real-time processing, our CAT-EDNet pays more attention to the salient detection accuracy while at the expense of increasing additional parameters and time cost. BASNet and UCtransnet both have superior performance, by contrast, our model has equal weight with BASNet while half size of transformer-based UCtransnet. In addition, our interference time can meet the real-time demand of actual manufacturing line. However, how to further compress the model and reduce the inference time is still in our future research work.

V. Conclusion

Incorporating defect integrity and defect boundary precision is a challenging task in salient detection of strip steel surface. In this paper, we propose a cross-attention transformer based encoder-decoder network (CAT-EDNet) to highlight the defect object and capture the fine boundary structure in the frequent noise occurred environment. The cross-attention transformer (CAT) with multi-head structure is embedded to the deeply supervised encoder-decoder like prediction module, and the aggregation weights of multi-scale layers are dynamically allocated to determine the salient region while considering the salient low-level details. In addition, the local-oriented cross-attention refinement module (CARM) is closely constructed to further optimize the boundary details in temporal domain. Extensive ablation studies have demonstrated the effectiveness of CAT and CARM in defect integrity and defect boundary precision. Compared with *twelve* state-of-the-art salient object detection methods on the noise randomly interfered SD-saliency-900 dataset, the *six* quantitative evaluation metrics, which are SM, WF, Mae, Eam, Dice and IoU, also prove the stronger noise robustness of our CAT-EDNet. Moreover, our model achieves real-time interference at a speed of 28.54 fps without any pre-preprocessing.

References

[1] Q. Luo, X. Fang, J. Su, J. Zhou, B. Zhou, C. Yang, *et al.*, "Automated Visual Defect Classification for Flat Steel Surface: A Survey," *IEEE Trans. Instrum. Meas.*, Oct. 2020, vol.69, no. 12, pp.9329-9349.

[2] Q. Luo, X. Fang, L. Liu, C. Yang, *et al.*, "Automated visual defect detection for flat steel surface: A survey," *IEEE Trans. Instrum. Meas.*, Feb. 2020, vol. 69, no. 3, pp. 626-644.

[3] Q. Luo, Y. Sun, P. Li, O. Simpson, L. Tian, and Y. He, "Generalized completed local binary patterns for time-efficient steel surface defect classification," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 3, pp. 667–679, Mar. 2019.

[4] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, " Frequency-tune salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1597–1604.

[5] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, and Q. Meng, "PGA-Net: pyramid feature fusion and global context attention network for automated surface defect detection," *IEEE Trans Ind. Informat.*, vol. 16, no. 12, pp. 7448-7458, Dec. 2020.

[6] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1493-1504, Apr. 2020.

[7] S. Ghorai, A. Mukherjee, M. Gangadaran, and P. K. Dutta, "Automatic defect detection on hot-rolled flat steel products," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 3, pp. 612–621, Mar. 2013.

[8] R. Achanta, S. Hemami, F. Estrada, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1597–1604.

[9] H. Peng, B. Li, H. Ling, *et al.*, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.

[10] Li, G, and Y. Yu, "Visual saliency based on multiscale deep features," *IEEE Conf. Comput. Vision Pattern Recogn.*, pp. 5455–5463, 2015.

[11] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7264–7273.

[12] Li Z, Lang C, Liew J H, *et al*. "Cross-layer feature pyramid network for salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 4587-4598, 2021.

[13] Qin X, Zhang Z, Huang C, *et al*. "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit*., 2020, pp. 107404.

[14] Lin S, Ryabtsev A, Sengupta S, *et al*. "Real-time high-resolution background matting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 8762-8771.

[15] Song, G., K. Song, and Y. Yan, "EDRNet: Encoder–Decoder Residual Network for Salient Object Detection of Strip Steel Surface Defects," *IEEE Trans. Instrum. Meas.*, vol. 1, no. 1, pp.99, 2020.

[16] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Proc. NIPS*, vol. 30, 2017, pp. 5998–6008.

[17] Liu, Yang, *et al.*, "A Survey of Visual Transformers," arXiv preprint arXiv:2111.06091, 2021.

[18] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc.CVPR*, 2021, pp. 6881–6890.

[19] Chen, J.; Lu, Y.; Yu, Q.; Luo, *et al.*, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv:2102.04306, 2021.

[20] E. Xie, W. Wang, Z. Yu, *et al.*, "Segformer: Simple and efficient design for semantic segmentation with transformers," arXiv:2105.15203, 2021.

[21] N. Carion, F. Massa, *et al.*, "End-to-end object detection with transformers," in *Proc. ECCV.* Springer, 2020, pp. 213–229.

[22] Y. Fang, S. Yang, X. Wang, *et al.*, "Queryinst: Parallelly supervised mask query for instance segmentation," arXiv:2105.01928, 2021.

[23] Peng C, Zhang X, Yu G, *et al.*, "Large kernel matters—improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1743–1751.

[24] Islam, Md Amirul, *et al.*, "Salient object detection using a context aware refinement network," In *BMVC*, 2017.

[25] Wang T, Borji A, Zhang L, *et al.*, "A stagewise refinement model for detecting salient objects in images," In *ICCV*, 2017, pp. 4039–4048.

[26] Zhang L, Dai J, Lu H, *et al.*, "A bidirectional message passing model for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2018, pp. 1741–1750.

[27] X. Qin, Z. Zhang, C. Huang, *et al.*, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7471–7481.

[28] Wang, Haonan, *et al.*, "UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-wise Perspective with Transformer," arXiv preprint arXiv:2109.04335, 2021.

[29] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," Ann. *Oper. Res*., vol. 134, no. 1, pp. 19–67, Feb. 2005.

[30] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *Proc. ISVC*, 2016, pp. 234–244.

[31] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput*., Pacific Grove, CA, USA, Nov. 2003, pp. 1398–1402.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in Proc. *IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1026–1034.

[33] T. Tieleman and G. Hinton, "RMSprop: Divide the gradient by a running average of its recent magnitude," *Neural Netw. Mach. Learn*., vol. 4, no. 2, pp. 26–31, 2012.

[34] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis*. (ICCV), Venice, Italy, Oct. 2017, pp. 4548–4557.

[35] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., Columbus, OH, USA, Jun. 2014, pp. 248–255.

[36] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., Providence, RI, USA, Jun. 2012, pp. 733–740.

[37] Fan, Deng-Ping, *et al.*, "Enhanced-alignment measure for binary foreground map evaluation.," arXiv preprint arXiv:1805.10421, 2018.

[38] Fidon, Lucas, *et al.*, "Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks," In *MICCAI-W*, pages 64–76, 2017.

[39] Nagendar, Gattigorla, *et al.*, "Neuro-IoU: Learning a Surrogate Loss for Semantic Segmentation," In *BMVC*. pp. 278, 2018.

[40] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., Salt Lake City, UT, USA, Jun. 2018, pp. 3089–3098.

[41] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*. (CVPR), Jun. 2019, pp. 3912–3921.

[42] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*. (CVPR), Jun. 2019, pp. 3902–3911.

[43] Zhao J, JJ Liu, "Egnet: Edge guidance network for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 8779–8788.

[44] Fan D P, Ji G P, "Camouflaged object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2020, pp. 2777–2787.

[45] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3080–3089.

[46] Chen, S., *et al.*, "Reverse Attention Based Residual Network for Salient Object Detection," *IEEE Trans. Image Process.*, vol. 1, no. 1, pp. 99, 2020.

QIWU LUO (M'17) received the B.S. degree in communication engineering from the National University of Defense Technology, Changsha, China, in 2008; and the M.Sc. degree in electronic science and technology and the Ph.D. degree in electrical engineering from Hunan University, Changsha, in 2011 and 2016, respectively.

He was a Senior Engineer of instrumentation with WASION Group Ltd. Company, Changsha, and the Deputy Technical Director with Hunan RAMON Technology Co., Ltd., Changsha. In 2016, he joined the School of Electrical Engineering and Automation, Hefei University of Technology, Hefei, China, where he also completed his postdoctoral research on automatic optic inspection (AOI). Since 2019, he has been an Associate Professor with the School of Automation, Central South University, Changsha. His current research interests include computer vision, industrial AOI, machine learning, parallel hardware architecture design, and reconfigurable computing.

JIAOJIAO SU received her B.S. degree in electronic information science and technology from the Hefei University of Technology in June 2020. She is currently pursuing a M.Sc. degree in advance in control science and engineering with the School of Automation, Central South University, Changsha, China, under the supervision of Dr. Luo. Her current research interests include defect detection, image classification, and machine learning.

CHUNHUA YANG (M'09) received the M.S. degree in automatic control engineering and the Ph.D. degree in control science and engineering from Central South University, Changsha, China, in 1988 and 2002, respectively.

From 1999 to 2001, she was a Visiting Professor with the University of Leuven, Leuven, Belgium. Since 1999, she has been a Full Professor with the School of Information Science and Engineering, Central South University, Changsha, China. From 2009 to 2010, she was a Senior Visiting Scholar with the University of Western Ontario, London, ON, Canada. She is currently the HoD of the School of Automation, Central South University. Her current research interests include modeling and optimal control of complex industrial processes, and intelligent control systems.

WEIHUA GUI (M'09) received the B.Eng. degree in electrical engineering from Central South University, Changsha, China, in 1976, respectively.

From 1986 to 1988, he was a visiting scholar with the University of Duisburg-Essen, Duisburg, Germany. Since 1991, he has been a Professor with the School of Automation, Central South University. Since 2013, he has been an Academician with the Chinese Academy of Engineering, Beijing, China. His research interests include modeling and optimal control of complex industrial processes, distributed robust control, and fault diagnosis.

LI LIU (M'08-SM'19) received the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), China, in 2012. She is currently a Professor with the College of System Engineering. During her PhD study, she spent more than two years as a Visiting Student at the University of Waterloo, Canada, from 2008 to 2010. From 2015 to 2016, she spent ten months visiting the Multimedia Laboratory at the Chinese University of Hong Kong. From 2016.12 to 2018.11, she worked as a senior researcher at the Machine Vision Group at the University of Oulu, Finland. Her current research interests include computer vision, pattern recognition and machine learning. Her papers have currently over 3700 citations in Google Scholar.

Olli Silvén received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Oulu, Finland, in 1982 and 1988, respectively. Since 1996, he has been a Professor of Signal Processing Engineering with the University of Oulu. His research interests include ultra-energy-efficient embedded signal processing and machine vision system design. He has contributed to the development of numerous solutions from real-time 3-D imaging in reverse vending machines to IP blocks for mobile video coding.