Interleaved Deep Artifacts-Aware Attention Mechanism for Concrete Structural Defect Classification

Gaurab Bhattacharya, Bappaditya Mandal, and Niladri B. Puhan, *Member, IEEE*

***Abstract*— Automatic machine classification of concrete structural defects in images poses significant challenges because of multitude of problems arising from the surface texture, such as presence of stains, holes, colors, poster remains, graffiti, marking and painting, along with uncontrolled weather conditions and illuminations. In this paper, we propose an interleaved deep artifacts-aware attention mechanism (iDAAM) to classify multi-target multi-class and single-class defects from structural defect images. Our novel architecture is composed of interleaved fine-grained dense modules (FGDM) and concurrent dual attention modules (CDAM) to extract local discriminative features from concrete defect images. FGDM helps to aggregate multi-layer robust information with wide range of scales to describe visually-similar overlapping defects. On the other hand, CDAM selects multiple representations of highly localized overlapping defect features and encodes the crucial spatial regions from discriminative channels to address variations in texture, viewing angle, shape and size of overlapping defect classes. Within iDAAM, FGDM and CDAM are interleaved to extract salient discriminative features from multiple scales by constructing an end-to-end trainable network without any preprocessing steps, making the process fully automatic. Experimental results and extensive ablation studies on three publicly available large concrete defect datasets show that our proposed approach outperforms the current state-of-the-art methodologies.**

***Index Terms*— Fine-grained dense module, concurrent dual attention module, concrete structural defect, convolutional neural network, multi-target multi-class classification.**

I. Introduction

THE rapid development of concrete infrastructures, such as bridges, highways, stadiums, tunnels, buildings and pavements attributes to the requirement of accurate, large-scale, automated and rapid inspection/monitoring methodologies. Otherwise, this leads to acceleration of the deterioration of damaged/unhealthy regions, raising potential threat of accidental collapse and large number of casualties (one example [1]). Image/video based inspection and monitoring of the unhealthy/defective regions involving automatic classification is very popular towards emerging and futuristic technological solutions for civil infrastructure management.

However, immense real-world challenges exist in obtaining automatic classification of defects because of the appearance of a wide variety of concrete surface textures as well as uncontrolled weather conditions, illumination, occlusion and capturing methodologies/devices. For damaged structures, this is exacerbated by inaccessibility and turbulent weather; making the visual inspection process dangerous, inaccurate, tedious and commonly surveyor-biased in nature [2]. The concrete defects come in large variants and are typically found in an overlapping manner, for example, an exposed bar defect might coexist with spallation and corrosion defects, which make the problem even more challenging for large concrete structures. The organisations managing civil infrastructures face immense challenges in maintaining the inspection along with their predictive analysis and monitoring of civil infrastructures. The task of inspection thus requires new innovative solutions incorporating computer vision and machine learning algorithms in conjunction with unmanned aerial vehicles (UAVs) to address overlapping defects with large unconstrained variations.

Over the last decade, deep convolutional neural networks (CNN) have been considered to achieve state-of-the-art performance for image classification using large multi-class datasets [3]–[9]. Inspired from this development, automatic concrete defect classification problem has become an active area of research over the last decade [10]–[14]. However, [10], [12] considered cracks as the only defect category in the images, thereby excluding all other important defect categories in structural health monitoring. Similarly, authors in [11], [13], [14] considered non-overlapping multi-class defects which do not address the real-world issue of having overlapping structural defects. Another common shortcoming to all these methods is that they are unable to apportion higher importance to the defective region from the healthy region within an image plane. Thereby, they process the entire image as a whole, emphasizing similar importance to both the defective and healthy regions of the image. The literature search dictates that the researchers have rarely worked with overlapping defect classes (such as spallation leading to exposed bar, which often leads to/co-exist with corrosion), which are often encountered in real-world applications with the challenging appearance variations, as discussed earlier. Recently, [16], [47] analyzed overlapping multi-class defects in CODEBRIM dataset using reinforcement learning and attention augmented CNN, respectively.

In this work, we aim to address the challenges mentioned for real-world concrete structural defect classification problems

Manuscript received August 28, 2020; revised March 23, 2021 and May 22, 2021; accepted July 13, 2021. Date of publication August 3, 2021; date of current version August 9, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christophoros Nikou. *(Corresponding author: Niladri B. Puhan.)*

Gaurab Bhattacharya and Niladri B. Puhan are with the School of Electrical Sciences, Indian Institute of Technology Bhubaneswar, Bhubaneswar 752050, India (e-mail: gb14@iitbbs.ac.in; nbpuhan@iitbbs.ac.in).

Bappaditya Mandal is with the School of Computing and Mathematics, Keele University, Newcastle ST5 5BG, U.K. (e-mail: b.mandal@keele.ac.uk).

Digital Object Identifier 10.1109/TIP.2021.3100556

using an interleaved artifacts-aware deep CNN architecture which effectively encapsulates the variations in degradation and unwanted inclusions. To accomplish this, a novel interleaved deep artifacts-aware attention mechanism (iDAAM) is proposed to classify both multi-target multi-class and single-class structural defect images. iDAAM architecture consists of interleaved fine-grained dense modules (FGDM) and concurrent dual attention modules (CDAM) to extract salient discriminative features from multiple scales to improve the classification performance. Experimental results and ablation studies show that the newly proposed architecture achieves significantly better classification performance than the state-of-the-art methodologies on three large datasets. Below we summarize our main technological contributions:

- We proposed fine-grained dense module to aggregate feature maps from multiple layers using identity mapping to obtain finer discriminative information from visually-similar overlapping concrete defect classes. The variation of relevant information can be captured by salient feature reuse at different layers which aids to efficient feature selection in subsequent attention modules.
- We proposed concurrent dual attention module that utilizes two new modules: firstly, a committee of multi-feature attention module to obtain multiple representations of highly localized features to address the overlapping defect classes occupying small regions and ensuring selection of more features using the parallel configuration to influence the classification performance. Secondly, the simultaneous excitation module which separately investigated the channel and spatial information to address the part deformation and shape variation present in overlapping defect classes. These two modules are aggregated to increase the ability of selecting relevant discriminative features among concrete defect classes.

  Finally, the fine-grained dense and concurrent dual attention modules are interleaved to obtain the iDAAM network, which does not require any preprocessing step and produces a focused feature selection mechanism on relevant defect regions.
- Ablation studies and extensive experimental evaluations for concrete defect classification on three state-of-the-art large datasets show the superiority of our iDAAM network over other state-of-the-art methods.

In the next Section, we describe the related work, Section III describes the proposed iDAAM architecture, Section IV and V present experimental results and ablation studies, respectively; Section VI provides the analysis and discussions on multiple datasets before drawing conclusions in Section VII.

II. Related Work

A. CNN for Image Classification

Large-scale image classification has become a pivotal problem in computer vision which observes exemplary success due to the deep CNN architectures such as AlexNet [3], VGG [4], GoogLeNet [5], etc. GoogLeNet incorporates multi-path feature extraction with filters of different kernel sizes, although considering large number of parameters [5]. ResNet [6] introduces skip connections from the previous layers to alleviate the vanishing gradient problem with a deeper architecture. Similarly, DenseNet [7] extends the skip connection addition to all previous layers to encapsulate larger variations in equivalent features. In our fine-grained dense module, we proposed to exploit feature reuse by creating identity mappings across multiple residual blocks to encapsulate defect variations with wide range of scales. The residual blocks inside this module performs feature extraction with filters of different kernel sizes, ensuring a deeper network with the ability to capture wide range of defect features. Also, the feature reuse strategy enables the network to aggregate fine-grained features with less parameters, as evident in Table V.

In literature, several multi-scale architectures such as Hourglass [49], feature pyramid network (FPN) [50], HR-Net [48], etc. are reported. In Hourglass architecture [49], multi-scale feature representation is obtained using the pooling layers and residual units involving large number of parameters compared to FGDM. In FPN [50], bottom-up top-down pathway is used to obtain multi-scale features by varying the spatial dimensions. In both Hourglass and FPN, each spatial plane is aggregated with the features having the same spatial resolution; however does not provide aggregation of responses across large range of scales, as provided by FGDM. On the other hand, HR-Net [48] uses multi-scale feature extraction and provides high resolution representations which are desirable for semantic segmentation, action recognition and pose estimation where precise spatial estimation is necessary. However, such high-resolution representations are not needed with the availability of small image patches for concrete defect classification.

B. Concrete Structural Defect Classification

Multiple research initiatives have been undertaken to classify concrete structural defects, although mostly involving single defect per image. Shi *et al.* [10] proposed a methodology using random structured forest to detect road cracks. Ye *et al.* considered the use of ZFNet architecture for touch panel glass surface having defects classes such as tilt, bubble and scratch [26]. In [20], [21], deep belief networks were incorporated for concrete defect prediction. Yang *et al.* performed single-target defect classification using AlexNet and VGG-based CNN models [11]. Dorafshan *et al.* [12] used deep CNN for classification of defects appearing in concrete bridge decks, walls and pavements (SDNET-2018). Multi-class single-target defect classification operation were performed using AlexNet and transfer learning in [13], [14]. Recently, Mundt *et al.* proposed reinforcement-learning based methodology for classification of five overlapping defect classes: crack, spallation, efflorescence, exposed bars, corrosion in the new CODEBRIM dataset [16].

C. Attention Mechanism

In recent years, attention mechanism with CNN framework has been proposed for extracting local discriminative features to achieve state-of-the-art performance. Initially, such

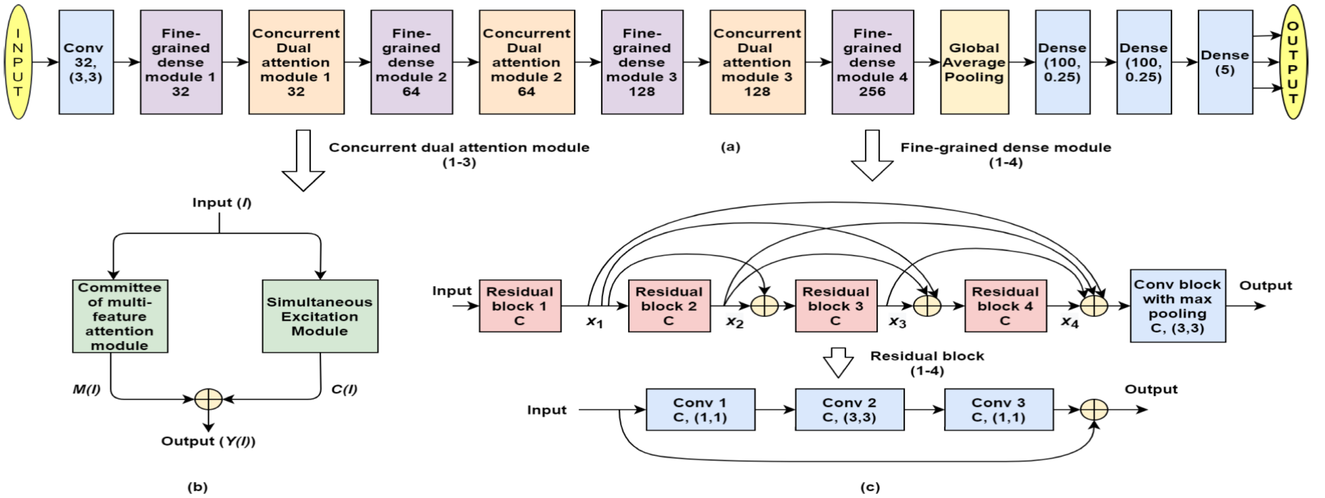
INPUT
Conv 32, (3,3)
Fine-grained dense module 1 32
Concurrent Dual attention module 1 32
Fine-grained dense module 2 64
Concurrent Dual attention module 2 64
Fine-grained dense module 3 128
Concurrent Dual attention module 3 128
Fine-grained dense module 4 256
Global Average Pooling
Dense (100, 0.25)
Dense (100, 0.25)
Dense (5)
OUTPUT
Concurrent dual attention module (1-3)
(a)
Fine-grained dense module (1-4)
Input (I)
Committee of multi-feature attention module
Simultaneous Excitation Module
M(I)
C(I)
Output (Y(I))
(b)
Input
Residual block 1 C
x_1
Residual block 2 C
x_2
Residual block 3 C
x_3
Residual block 4 C
x_4
Conv block with max pooling C, (3,3)
Output
Residual block (1-4)
Input
Conv 1 C, (1,1)
Conv 2 C, (3,3)
Conv 3 C, (1,1)
Output
(c)

Fig. 1. (a) Architecture of iDAAM. Here Conv block represents convolutional operation with first number representing the number of filters and the next two numbers give the filter dimension for each channel. Dense represents the dense layer, where first number gives the number of nodes and the second number is the dropout value. (b) Proposed concurrent dual attention module composed of committee of multi-feature attention module and simultaneous excitation module. The number denoted in (a) with the concurrent dual attention module represents filter size, explained in Fig. 2(a) and Fig. 2(c). (c) Proposed fine-grained dense module. The numbers denoted in (a) with the fine-grained dense module represent number of filters for convolution. The max pooling operation uses pool size of (2,2) and stride 2.

networks are used for analysing sequential data [17] and also for general image classification [19]. Park *et al.* [22] and Woo *et al.* [23] investigated the impact of channel and spatial attention modules for discrimination of features. Attention modules have been applied for object detection [24], [25], multi-label classification [57], action recognition [27]–[29], image captioning [30]–[32], re-identification [33], [34], saliency prediction [35], pedestrian attribute recognition [36], etc.

Use of visual attention in concrete defect classification is however limited; except a recent work [47] involving residual attention mechanism. However, the considerable amount of parameters and computations involved in parallel feature extraction make the network inferior for real-time applications. Contrary to only using residual attention, in our iDAAM architecture, we have proposed concurrent dual attention module, which consolidates several types of attention for better discrimination. The committee of multi-feature attention module in our network aggregates multiple feature representation, rather than obtaining a single representation as in [17]. Similarly, we perform spatial squeezing operation inside simultaneous excitation module using 1 × 1 convolution, rather than using the pooling operation in [22], [23] to adaptively generate local spatial descriptors. Moreover, the iDAAM architecture attributes to significant reduction in the number of parameters compared to the state-of-the-art [47] due to the absence of explicit multi-branch feature extraction.

III. Proposed Methodology

The proposed iDAAM architecture is composed of interleaved fine-grained dense modules (FGDM) and concurrent dual attention modules (CDAM), which jointly helps to extract salient discriminative features from multiple scales to improve both multi-target multi-class and single-class classification performance. The structure of the proposed iDAAM architecture is shown in Fig. 1, and we describe these modules in detail and summarize their roles at the end of their respective subsections.

A. Fine-Grained Dense Module

To obtain robust discriminative features from similar-looking defect classes such as spallation and efflorescence within a small region in the image, we need finer information which can be aggregated by employing a deep network. However, experimental results suggest that in a deep network, the weights in a particular layer could not update due to the small value of the gradient. This is known as the “vanishing gradient problem” and it results in the downfall of performance. This problem can be alleviated by using short paths from the former layer to later layer using the residual connection without extra parameters and computations [6]. Different from [6], we propose fine-grained dense module which enhances the potential of residual connections by exploiting feature reuse using identity mappings across multiple residual blocks to capture the variations present in concrete defects. Such aggregation of salient features from similar-looking defect classes at different layers makes the relevant feature selection of subsequent attention modules easy and efficient, as shown in Fig. 1.

In the fine-grained dense module, the output of L^{th} residual block is added with the feature-maps extracted by all previous residual blocks, i.e. $x_1, \ldots, x_{L-1}$, resulting in the final response of L^{th} residual block x_L.

$$\begin{aligned} x_1 &= H_1(I), \\ x_L &= H_L(x_1 + x_2 + \ldots + x_{L-1}), \quad L \in [2, 4]. \end{aligned} \tag{1}$$

Here, H_L is the function performed by the residual block and I is the input to this module. The proposed fine-grained

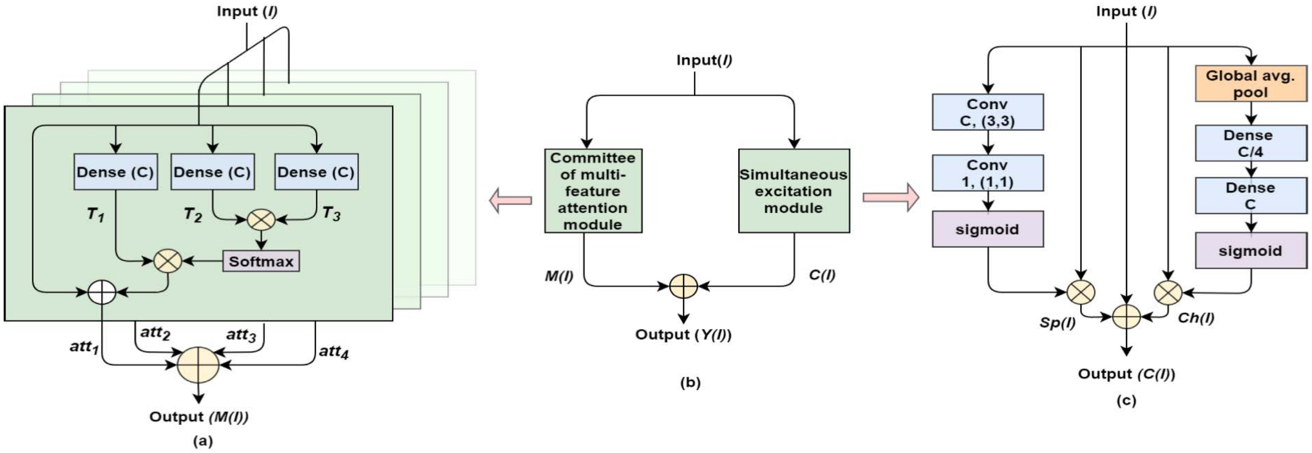

Fig. 2. (a) Proposed committee of multi-feature attention module. (b) Concurrent dual attention module. (c) Simultaneous excitation module.

dense module is designed by interconnecting four residual blocks, where each residual block is considered to be a three-layer convolutional network with the identity mapping as shown in Fig. 1 (c). The residual block operation,

$$Res(I) = Conv(Conv(Conv(I))) + I. \tag{2}$$

Finally, the outcome of the dense interconnection passes through a *Conv* layer followed by max pooling operation to reduce computations in extracting finer features. Overall, fine-grained dense module helps to aggregate multi-layer robust information to describe visually-similar overlapping defects.

B. Concurrent Dual Attention Module

The novel concurrent dual attention module incorporates multiple attention operations (self, spatial and channel attention) in parallel manner and aggregates their outcomes to make discriminative features more prominent for multi-target defect classes [17], [22], [23]. Fig. 1 (b) describes the block diagram, which consists of (a) committee of multi-feature attention module, and (b) simultaneous excitation module. Committee of multi-feature attention module has been designed to encode multiple representations of highly localized features which enable the network to learn minute defect classes. The spatial and channel information for artifacts are encoded in a parallel manner using the simultaneous excitation module which concurrently highlights the relevant features and reduces the impact of weak or unimportant features. The outputs of these two attention modules are aggregated to achieve greater impact for the concurrent dual attention module.

1) Committee of Multi-Feature Attention Module: To encode salient information from visually-similar, overlapping variable-sized concrete structural defects, we exploit multiple representations of highly localized parallel feature extraction mechanism using committee of multi-feature attention module. This module helps to encapsulate highly localized feature selection mechanism to distinguish between concrete defects and various uncontrolled artifacts such as graffiti, poster remains, small holes, etc. The parallel configuration enables the detection of overlapping defect classes (such as corrosion, efflorescence and spallation), which usually occupy very small regions to influence the classification performance.

In this proposed module, the attention operations are performed multiple times to ensure that maximum important features are attended. Each attention module performs parallel operations using three dense layers to obtain parallel non-linear projections in feature space. Here the input I is considered with height, width and number of channels as H, W, C, respectively.

$$I(H, W, C) \rightarrow Dense(C) \rightarrow T_i(H, W, C), \quad \forall i \in [1, 3]. \tag{3}$$

Thereafter, the outputs T_2 and T_3 are multiplied element-wise, passed through a softmax function to generate the attention mask, and is then multiplied with T_1 to highlight the important features. Next, the identity mapping is performed by the addition of input tensor to the output.

$$att_k = T_1 \times softmax(T_2 \times T_3) + I(H, W, C), \quad k \in [1, 4]. \tag{4}$$

Finally, all the four outputs obtained by attention operations att_1, att_2, att_3 and att_4 are added; giving the output of the CMFA module which aggregates attentive features from multiple representations.

$$M(I) = att_1 + att_2 + att_3 + att_4. \tag{5}$$

2) Simultaneous Excitation Module: The convolution layer captures local spatial features across all the channels and thus jointly encodes relevant spatial and channel information [6], [42]. However, for our case, the overlapping concrete defect classes attribute to viewing angle variations and part deformation due to variation in texture and shape of structures. Also, we need to selectively highlight the channel-wise discriminative defect features while suppressing others [22], [23], [42]. To address these problems, simultaneous excitation module separately investigates the relevant spatial and channel information to improve the performance.

One part of this proposed module performs the squeezing of the spatial plane of the input tensor using global average pooling and then exciting it channel-wise to obtain channel

TABLE I

COMPARISON OF MULTI-TARGET VALIDATION ACCURACY (%) AND BEST VALIDATION MODEL'S MULTI-TARGET TEST ACCURACY (%) BY VARYING THE INPUT IMAGE SIZES AND BATCH SIZES FOR iDAAM ARCHITECTURE USING CODEBRIM DATASET

Input image size	**Batch size: 16**			**Batch size: 32**		
	Train accuracy	**Val. accuracy**	**Test accuracy**	**Train accuracy**	**Val. accuracy**	**Test accuracy**
96	**99.98**	**91.82**	**89.54**	99.93	89.94	88.46
128	99.94	90.93	88.23	99.89	90.52	88.16
160	99.87	89.74	88.04	99.83	89.95	87.84
192	99.49	88.35	87.72	99.37	88.05	86.93

TABLE II

COMPARISON OF THE CLASSIFICATION ACCURACY (%) OF iDAAM ARCHITECTURE WITH THE STATE-OF-THE-ART METHODS ON CODEBRIM DATASET

Architecture	**Multi-target accuracy**		**Parameters in million**
	Best validation	**Best val-test**	
AlexNet [3]	63.05	66.98	57.02
T-CNN [8]	64.30	67.93	58.60
VGG-A [4]	64.93	70.45	128.79
VGG-D [4]	64.00	70.61	134.28
WRN-28-4 [9]	52.51	57.19	5.84
Densenet-121 [7]	65.56	70.77	11.50
SE-ResNet-50 [42]	72.86	70.71	28.13
ResNeSt [43]	75.92	73.46	27.50
ENAS-1 [40]	65.47	70.78	3.41
ENAS-2 [40]	64.53	68.91	2.71
ENAS-3 [40]	64.38	68.75	1.70
MetaQNN-1 [41]	66.02	68.56	4.53
MetaQNN-2 [41]	65.20	67.45	1.22
MetaQNN-3 [41]	64.93	72.19	2.88
MDAL [47]	86.15	84.29	10.43
iDAAM	**91.82**	**89.54**	**4.89**

information. The squeezing operation across the channels enables the module to implicitly embed the global channel description, providing channel-wise statistics of the entire image. The following dense layers exploit contextual channel information with non-linear adaptive re-calibration and their inter-relationship helps to extract discriminative channels with crucial features. The channel information is multiplied with the input for highlighting the features relevant for discrimination along the channels, *Ch(I)*. In channel attention, if we consider the input to be *I(H, W, C)* $= [I_1, I_2, \ldots, I_C]$, then the output of this operation *U(C)* can be written as in (6).

$$U(k) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} I_k(i, j), \quad \forall k \in [1, C]. \tag{6}$$

Hence by this operation, the global spatial information from I is embedded in U, which is further connected to the dense layers. Finally, the outcome of two layers are activated using sigmoid function and then multiplied with the input to obtain channel attention feature map *Ch(I)*.

$$Ch(I) = sigmoid(W_1 \times (ReLU(W_2 \times U))) \times I. \tag{7}$$

Here W_1 and W_2 represent the weights used in the two dense layers to generate channel attention.

Similarly, other part of the simultaneous excitation module squeezes the channels using the convolution blocks which capture the spatial features predominant across all channels. The extracted features are spatially excited and then the output is multiplied with the input tensor to give prominence to relevant spatial information *Sp(I)*.

$$Sp(I) = sigmoid(Conv(ReLU(Conv(I)))) \times I. \tag{8}$$

Unlike [23], where the spatial attention was performed using average and max pooling operation; in our proposed module, the global channel features are squeezed to extract relevant spatial information to generate spatial statistics by shrinking the input through its channel dimension. We anticipate that the proposed usage of 1 × 1 convolution across all channels can be interpreted as a collection of local spatial descriptors. To aid to the efficient feature extraction, another convolution block is added before the application of 1 × 1 convolution for aggregation of spatial information, rather than using directly for feature aggregation. Furthermore in our case, the input is added using skip connection to avoid missing important discriminative cues and alleviate the vanishing gradient problem. Thereby, the total response $C(I)$ is given by:

$$C(I) = Sp(I) + Ch(I) + I. \tag{9}$$

Finally, we add the outputs $M(I)$ and $C(I)$ to aggregate highly localized parallel features from committee of multi-feature attention module and spatial and channel information from simultaneous excitation module. The output for the concurrent dual attention module $Y(I)$ is given by:

$$Y(I) = M(I) + C(I). \tag{10}$$

To increase the receptivity and chance of getting relevant local discriminative features, the concurrent dual attention modules are stacked with fine-grained dense modules. We hypothesize that it is difficult for a single attention module to extract all discriminative local features from the complex multi-target multi-class structural images.

IV. EXPERIMENTAL RESULTS

The performance of the iDAAM architecture for concrete defect classification is evaluated using three large structural defect image datasets [12], [16], [37].

A. Results on CODEBRIM Dataset

We conduct our experiments on the current state-of-the-art and most challenging dataset with overlapping defects: COncrete DEfect BRidge IMage (CODEBRIM) [16], obtained for non-commercial research and educational purpose. This dataset was prepared using multi-target multi-class concrete defect images with varying ranges of illumination, humidity and resolution by investigating 30 bridges with different degrees of deterioration, surface roughness and weather condition to capture the possible changes for real-world application. A total of 5354 annotated overlapping defect images and 2506 background images are generated, which include 2507 images with crack, 1898 images with spallation,

TABLE III

COMPARISON OF CRACK DEFECT CLASSIFICATION ACCURACY (%) OF iDAAM ARCHITECTURE WITH THE STATE-OF-THE-ART METHODS FOR CONCRETE BRIDGE DECK, WALL AND PAVEMENT ON SDNET-2018 DATASET

Model Description	Bridge image result			Wall image result			Pavement image result		
	Train accuracy	Val. accuracy	Test accuracy	Train accuracy	Val. accuracy	Test accuracy	Train accuracy	Val. accuracy	Test accuracy
Alexnet, fully trained [12]	98.25	94.43	91.86	97.52	90.26	87.88	98.48	97.15	95.22
Alexnet, Transfer learning [12]	98.78	95.84	92.07	98.34	92.59	90.16	99.06	97.54	95.85
VGG16, without image augmentation [39]	98.55	86.45	85.19	97.25	88.24	84.29	99.14	89.79	88.56
VGG16 with augmentation [39]	96.35	90.15	87.76	94.55	91.24	86.29	97.59	94.37	89.33
VGG16 with augmentation and transfer learning [39]	94.22	90.25	88.59	93.89	91.86	87.46	97.58	93.45	92.13
VGG16 with transfer learning, augmentation,fine tuning [39]	98.59	94.36	92.79	97.28	93.88	91.48	99.12	97.59	96.78
Inception [5]	98.76	94.58	92.86	97.58	94.83	92.75	99.32	97.89	97.31
ResNet-50 [6]	98.49	95.88	93.15	97.96	95.08	92.36	99.15	98.11	97.28
Densenet-121 [7]	98.85	96.03	93.58	98.12	97.49	93.19	99.46	98.27	97.59
SE-ResNet-50 [42]	98.96	96.25	94.18	98.36	97.58	93.79	99.32	98.29	97.36
ResNeSt [43]	99.03	96.32	93.96	98.41	97.46	94.22	99.19	98.35	97.61
MDAL [47]	99.91	98.56	94.35	98.79	98.12	93.76	99.94	98.92	98.26
iDAAM	**99.15**	**97.23**	**95.38**	**98.92**	**96.72**	**95.16**	**99.48**	**98.76**	**98.12**

TABLE IV

COMPARISON OF THE CLASSIFICATION ACCURACY (%) OF THE iDAAM ARCHITECTURE WITH THE STATE-OF-THE-ART METHODS ON CONCRETE CRACK IMAGE DATASET

Model name	Training accuracy(%)	validation accuracy (%)	Testing accuracy (%)
VGG [4]	97.25	97.00	96.80
Inception [5]	98.00	97.85	97.60
Resnet-50 with transfer learning [38]	98.40	98.00	97.80
Deep CNN with adaptive threshold [15]	99.75	99.16	98.70
DenseNet-121 [7]	98.85	98.45	98.00
SE-ResNet-50 [42]	99.75	99.70	99.60
ResNeSt [43]	99.80	99.65	99.55
MDAL [47]	99.99	99.84	99.81
iDAAM	**99.98**	**99.84**	**99.78**

833 images with defect as efflorescence, 1507 images with exposed bars and 1559 images with corrosion stain.

The validation and test images are selected by following the implementation protocol in [16], where we perform "precise estimate of a model's performance in a multi-target scenario, a classification is considered as correct if, and only if, all the targets are predicted correctly". Also, this dataset provides image sizes with large variations. Hence, the network is trained using different input image and mini-batch sizes for 200 epochs with learning rate 0.001 and momentum 0.9. To perform multi-target multi-class classification, we use sigmoid activation function for every class in the final dense layer and the loss function is considered to be binary cross-entropy. From Table I, it is evident that the iDAAM network achieves its highest test accuracy of 89.54% with the input image dimension of 96 × 96 × 3 and batch size of 16.

The convergence curves are generated during the training of iDAAM network considering image dimensions 96 & 128 with mini-batch sizes 16 & 32 and are shown in Fig. 3. It can be observed that image dimension 96 achieves faster and more stable convergence when trained using mini-batch size 16 and also obtains better classification accuracy; hence indicating the best possible choice to investigate the network for CODEBRIM dataset.

Table II shows the comparative performance analysis of the proposed iDAAM architecture with the state-of-the-art methods, as reported in [16]. It is evident from the table that the iDAAM architecture outperforms all the existing state-of-the-art methods by a significant amount, i.e. 89.54% compared to 84.29% given by MDAL [47]. The CODEBRIM dataset contains diverse ranges of images with variations in scale, aspect ratio, resolution and uncontrolled artifacts and hence the training, validation and test subsets have large variations in defect deformation which impact the test performance in spite of having very high training accuracy. However, from Table II, we can observe that the efficient utilization of attention modules interleaved within the iDAAM architecture helps to alleviate these problems and thus significantly outperforms the current state-of-the-art methodologies.

B. Results on Concrete Crack Image Dataset

This dataset is a collection of 40000 images of 20000 crack and 20000 non-crack images of dimension $227 \times 227 \times 3$ [37], obtained under a Creative Commons Attribution 4.0 International license. To evaluate the iDAAM architecture, 32000 random images are considered for training, 4000 for validation and rest 4000 for testing with equal numbers of crack and non-crack images in every subsets, following the protocol in [38]. The network is trained with batch size 16, learning rate 0.001 and momentum 0.9 for 200 epochs. To make the iDAAM architecture suitable for concrete crack image and SDNET-2018 datasets for single-target crack defect classification, two nodes are used in the output classification layer with softmax activation and categorical cross-entropy loss function is used for training purpose. Table IV shows the performance of the iDAAM for this dataset which outperforms many state-of-the-art methods and gives very similar performance of [47], achieving the performance with 2.13× less parameters than [47]. The convergence curve of iDAAM using this dataset is shown in Fig. 3 (e). From this curve, we can observe that the network converges faster which is mainly due to the less variations of concrete crack image dataset than CODEBRIM dataset and the presence of only crack defects, which helps to

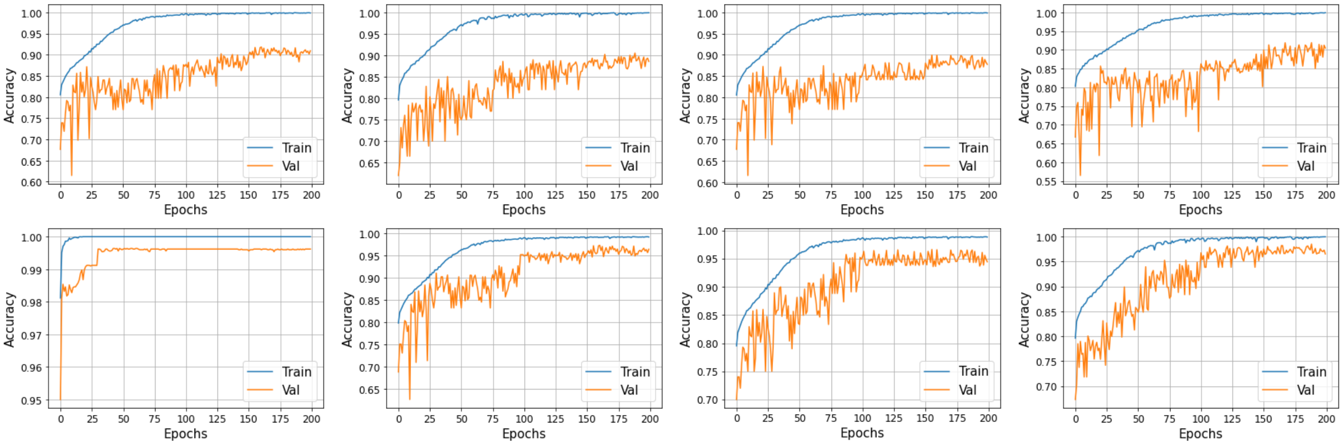

Fig. 3. Convergence curves during training using the iDAAM architecture on CODEBRIM, concrete crack image and SDNET-2018 datasets. Here blue and orange curves represent training data and validation data accuracy over the epochs, respectively. Top row, from left to right (CODEBRIM): (a) Curves for batch size 16 and image dimension 96, (b) Curves for batch size 16 and image dimension 128, (c) Curves for batch size 32 and image dimension 96, (d) Curves for batch size 32 and image dimension 128. Bottom row, from left to right: (e) Curves for concrete crack image dataset, (f) Curves for bridge deck images from SDNET-2018, (g) Curves for wall images from SDNET-2018, (h) Curves for pavement images from SDNET-2018.

achieve very high testing accuracy due to the robust feature selection mechanism.

C. Results on SDNET-2018 Dataset

SDNET-2018 dataset was presented by Dorafshan *et al.* [12], obtained under Attribution 4.0 International licensing. This dataset is made up of more than 56000 annotated crack and non-crack images of dimension $256 \times 256 \times 3$. These image patches are segmented from 230 images of different concrete structures (72 wall images, 54 bridge deck images and 104 pavement images) with varying crack widths and obstacles such as clutter, illumination, shadows and unwanted inclusions.

In our work, for a fair comparison with other methods, we have followed the implementation protocol given in [12]. During the execution, the dataset split is performed and for each training, 200 epochs are considered with mini-batch size of 16 and learning rate 0.001 and momentum 0.9. From the experimental results shown in Table III, it is evident that our proposed iDAAM architecture achieves superior performance as compared to the previous methods for bridge and wall images with higher accuracy. For the pavement images, the performance is very similar to [47] with 2.13× reduction in parameters. The convergence curves of iDAAM architecture for the bridge deck, wall and pavement image subsets from SDNET-2018 dataset are depicted in Fig. 3 (f), (g) and (h), respectively, which demonstrate that the network converges slower than the concrete crack image dataset, due to the variation in surface textures and uncontrolled artifacts present in SDNET-2018 dataset. However, they converge faster than the network in case of CODEBRIM dataset due to the presence of single crack defects and no variations in scale, resolution and aspect ratio as found in CODEBRIM dataset.

For iDAAM architecture, the average training time in each epoch for batch size 16 is 67.2 seconds and average testing time per image is 0.135 ms on CODEBRIM dataset. For the implementation, we have used Python keras 2.3.1 api with Tensorflow 1.13.2 in the backend on a system with Intel Core i7 processor, 16 GB RAM, and NVIDIA GeForce RTX-2070 8GB GPU card. The codes for the proposed modules can be accessed in [58].

V. Ablation Studies

A series of ablation study experiments are carried out on three datasets to understand the importance of fine-grained dense modules and concurrent dual attention modules. Due to the vast nature of the ablation analysis, we have divided the experiments in two parts, one for the fine-grained dense module and another for the concurrent dual attention module. In Tables V, VI, VII and VIII and in the following discussion, FGDM stands for fine-grained dense module, CDAM stands for concurrent dual attention module, CMFA represents committee of multi-feature attention and SEM means simultaneous excitation module.

A. Ablation Experiments on Fine-Grained Dense Module

We perform an extensive set of experiments on FGDM to analyze its impact under various alterations and report the results on CODEBRIM and concrete crack defect datasets in Table V and Table VI. Firstly, the proposed network is trained without using any fine-grained dense module, i.e. only considering the concurrent dual attention modules, which consumes less parameters (1.58 million). However, the absence of the fine-grained dense module leads to poor classification performance due to the unavailability of fine-grained information, as evident in Tables V and VI. Moreover, we can observe a significant drop in performance for CODEBRIM dataset due to its complex nature and large variations in appearance.

Secondly, to show the generality of our proposed FGDM, we replace the residual block present in FGDM with inception module [5], which increase the number of parameters to 19.65 millions as compared to 4.89 millions in iDAAM, although producing reduced classification performance.

TABLE V

ABLATION STUDIES ON FINE-GRAINED DENSE MODULES ON THE CLASSIFICATION ACCURACY(%) IN THE iDAAM ARCHITECTURE ON CODEBRIM AND CONCRETE CRACK IMAGE DATASET

Model Description	Parameters in million	CODEBRIM Dataset			Concrete Crack Image dataset		
		Training accuracy	Val. accuracy	Test accuracy	Training accuracy	Val. accuracy	Test accuracy
iDAAM without FGDM	1.58	97.45	80.45	73.89	99.14	98.27	97.88
CDAM + Inception within FGDM	19.65	98.75	87.56	86.14	99.97	99.75	99.18
Res2Net [18] replacing FGDM	12.35	99.25	89.87	88.38	99.97	99.79	99.25
HRNet [48] replacing FGDM	11.95	99.48	90.84	88.12	99.97	99.75	99.45
Hourglass [49] replacing FGDM	16.29	99.39	89.25	87.96	99.97	99.65	99.15
FPN [50] replacing FGDM	10.56	99.26	88.98	87.59	99.96	99.45	99.10
iDAAM	**4.89**	**99.98**	**91.82**	**89.54**	**99.98**	**99.84**	**99.78**

TABLE VI

ABLATION STUDIES ON FINE-GRAINED DENSE MODULES ON THE CLASSIFICATION ACCURACY(%) IN THE iDAAM ARCHITECTURE ON CONCRETE BRIDGE DECK, WALL AND PAVEMENT FROM SDNET-2018 DATASET

Model Description	Bridge image result			Wall image result			Pavement image result		
	Training accuracy	Val. accuracy	Test accuracy	Training accuracy	Val. accuracy	Test accuracy	Training accuracy	Val. accuracy	Test accuracy
iDAAM without FGDM	94.81	93.11	91.25	95.68	92.79	90.57	97.62	95.76	93.89
CDAM + Inception within FGDM	98.67	96.67	94.29	98.61	96.17	94.23	99.27	98.28	96.94
Res2Net [18] replacing FGDM	98.92	96.85	94.89	98.82	96.35	95.02	99.38	98.46	97.19
HRNet [48] replacing FGDM	98.76	96.91	94.85	98.49	96.26	95.11	99.27	98.46	97.09
Hourglass [49] replacing FGDM	98.21	96.34	94.16	98.29	96.06	94.89	99.11	98.20	96.75
FPN [50] replacing FGDM	98.03	95.86	93.49	98.13	95.85	94.38	98.86	97.25	96.40
iDAAM	**99.15**	**97.23**	**95.38**	**98.92**	**96.72**	**95.16**	**99.48**	**98.76**	**98.12**

Thirdly, we experimented the impact of FGDM by replacing it with the Res2Net module [18]. The FGDM considers the entire input feature maps to undergo feature extraction operation in residual blocks to extract the local features across all the channels, resulting in a deeper network. On the other hand, the Res2Net module divides the entire feature map into mutually exclusive sub-parts, where each part undergoes feature extraction operation, resulting in a granular network. In FGDM, each residual block aggregates features across all previous responses to capture multi-scale feature representation. However, in Res2Net, each part aggregates features only from the previous part of the input instead of capturing all previous responses. Furthermore, in FGDM, multi-scale feature representation is performed using the identity mapping, which does not involve extra parameters. In Res2Net, multi-scale feature representation is performed using the 3×3 convolution operation. The result shows that by replacing the FGDM module with Res2Net, we obtain reduced performance (88.38% test accuracy compared to 89.54% by iDAAM on CODEBRIM) with large increment in the number of parameters (12.35 million compared to 4.89 million); which shows the efficacy of a deeper network such as FGDM to obtain robust features for overlapping defect recognition. Similarly, we extend the same operation by replacing FGDM with several state-of-the-art multi-scale feature extraction units, such as HR-Net [48], Hourglass [49] and FPN [50]. Unlike these modules, FGDM focuses on long-range representation and the interconnected residual blocks further fine-tune the features. Also, these networks focus on precise spatial feature estimation; however for structural defect classification, we require rich long-range discriminatory information. Moreover, the identity mappings across the residual units enable the network to alleviate vanishing gradient problem with generalized performance. These characteristics of FGDM help in the improved performance for all three concrete defect datasets.

B. Ablation Experiments on Concurrent Dual Attention Module

Similar to FGDM, we evaluate various aspects of CDAM with an extensive ablation study and report the results on CODEBRIM and concrete crack image datasets in Table VII and SDNET-2018 in Table VIII. Firstly, we train the network keeping only FGDM, i.e. removing all CDAMs. Here the performance drop is experienced due to the absence of attentive feature extraction mechanism. Similar to the previous result, we can observe worse performance for the CODEBRIM dataset due to large variations and presence of uncontrolled artifacts, which could have been properly discriminated using the entire attention mechanism.

Secondly, we try to analyze the importance of different attention operations present in the iDAAM architecture. To investigate this, the SEMs are removed, while keeping all the FGDMs. Experimental results (in Tables VII and VIII) illustrate drop in performance for CODEBRIM and SDNET-2018 datasets due to the absence of spatial-channel attention mechanism to encode deformed defects and texture variation. Then the same operation is replicated by removing all the CMFA modules and keeping others same, which shows that the CODEBRIM dataset produces less accuracy due to the absence of highly localized feature selection required to discriminate between similar-looking overlapping defects.

Thirdly, we remove one of the channel and spatial attention parts of the SEM sub-network at a time keeping other modules intact to analyze the impact of individual attention operations

TABLE VII

ABLATION STUDIES ON CONCURRENT DUAL ATTENTION MODULES ON THE CLASSIFICATION ACCURACY(%) IN THE iDAAM ARCHITECTURE ON CODEBRIM AND CONCRETE CRACK IMAGE DATASET

Model Description	**Parameters in million**	**CODEBRIM Dataset**			**Concrete Crack Image dataset**		
		Training accuracy	**Val. accuracy**	**Test accuracy**	**Training accuracy**	**Val. accuracy**	**Test accuracy**
iDAAM without CDAM	4.61	94.55	78.24	69.74	98.26	97.89	97.25
FGDM + Only CMFA	4.87	99.26	89.29	83.56	99.96	99.73	99.21
FGDM + Only SEM	4.63	99.15	90.11	84.27	99.97	99.69	99.17
Only spatial attention in SEM	4.88	99.89	91.35	88.93	99.97	99.75	99.70
Only channel attention in SEM	4.88	99.92	91.68	89.01	99.97	99.81	99.75
CMFA with one layer	4.85	99.12	88.56	82.98	99.95	99.45	99.25
CMFA with dot product operation	4.89	99.75	91.12	88.58	99.97	99.65	99.50
CMFA with concatenation operation	4.93	99.89	91.56	89.21	99.98	99.75	99.50
h-SE-ResNet [44] replacing channel attention	5.14	99.94	91.68	88.58	99.98	99.80	99.76
CI-BCNN [45] replacing channel attention	4.91	99.91	90.88	87.63	99.97	99.70	99.60
CSAR [46] replacing channel attention	4.93	99.93	90.78	88.46	99.96	99.65	99.55
iDAAM with 2x channels	18.76	99.51	91.35	88.38	99.97	99.65	99.35
Self-attention replacing CMFA	4.93	99.26	87.46	85.11	99.97	99.75	99.25
Self-attention replacing SEM	4.93	99.53	90.12	84.23	99.96	99.78	99.32
Self-attention replacing CDAM	4.62	95.16	81.25	75.86	98.35	98.20	97.85
iDAAM	**4.89**	**99.98**	**91.82**	**89.54**	**99.98**	**99.84**	**99.78**

TABLE VIII

ABLATION STUDIES ON CONCURRENT DUAL ATTENTION MODULES ON THE CLASSIFICATION ACCURACY(%) IN THE iDAAM ARCHITECTURE ON CONCRETE BRIDGE DECK, WALL AND PAVEMENT FROM SDNET-2018 DATASET

Model Description	**Bridge image result**			**Wall image result**			**Pavement image result**		
	Training accuracy	**Val. accuracy**	**Test accuracy**	**Training accuracy**	**Val. accuracy**	**Test accuracy**	**Training accuracy**	**Val. accuracy**	**Test accuracy**
iDAAM without CDAM	94.63	92.58	91.06	95.79	93.04	90.24	97.88	94.89	93.56
FGDM + Only CMFA	98.64	95.59	93.76	97.52	94.29	93.86	98.67	97.53	97.08
FGDM + Only SEM	98.59	96.48	92.89	97.18	94.56	92.49	98.92	97.66	95.48
Only spatial attention in SEM	98.92	97.05	95.16	98.46	96.35	94.92	99.13	98.27	97.68
Only channel attention in SEM	99.08	97.14	95.21	98.55	96.41	95.00	99.31	98.32	97.93
CMFA with one layer	98.75	96.42	93.89	98.26	95.78	94.69	98.37	97.21	96.18
CMFA with dot product operation	98.89	96.97	94.74	98.35	96.21	94.89	99.05	98.17	97.56
CMFA with concatenation operation	98.95	97.13	94.98	98.58	96.67	95.02	99.23	98.49	97.86
h-SE-ResNet [44] replacing channel attention	98.91	97.08	94.81	98.46	96.55	95.03	99.19	98.46	97.88
CI-BCNN [45] replacing channel attention	98.46	96.83	94.31	98.05	96.12	94.75	98.96	98.19	97.49
CSAR [46] replacing channel attention	98.86	96.87	94.58	98.34	96.27	94.79	99.12	98.30	97.69
iDAAM with 2x channels	99.06	96.89	95.12	98.53	96.25	94.76	99.26	98.24	97.49
Self-attention replacing CMFA	98.74	96.12	93.27	97.31	95.27	93.66	99.15	97.83	96.42
Self-attention replacing SEM	98.76	96.12	94.09	97.84	95.27	94.21	98.91	97.83	97.39
Self-attention replacing CDAM	95.49	93.26	91.78	96.38	93.27	91.42	98.05	95.17	94.39
iDAAM	**99.15**	**97.23**	**95.38**	**98.92**	**96.72**	**95.16**	**99.48**	**98.76**	**98.12**

in the recognition performance. Tables VII and VIII give the concrete defect recognition performance by dropping one of the attention sub-networks which demonstrates reduced performance in both cases, further re-instating the reason for using both channel and spatial attention parts.

Fourthly, we perform multiple experiments to obtain design justification of CMFA sub-network. For this, we first conduct an experiment by replacing the CMFA module containing four layers with only one such layer. The experimental results depict a decline in performance due to the absence of aggregation of multiple discriminative feature representation. However, incorporation of five such layers in CMFA gives similar performance with added parameters and hence is not reported. Then, we alter the additive operation in the CMFA modules with the dot product and concatenation operation to obtain the best aggregation method to be used for the attention. From the results in Tables VII and VIII, we observe that the use of dot product reduces the testing accuracy on CODEBRIM (88.58% compared to 89.54%) without change in the number of parameters, whereas the use of concatenation operation gives similar performance (89.21% compared to 89.54%), but with extra parameters required in the dense layers for linear projection. Hence, we infer that addition attention is most suited for this application.

Furthermore, we perform a series of experiments to realize the impact of channel attention part present in SEM. For this, we first replace the channel attention part in SEM with several state-of-the-art channel attention mechanisms such as hierarchical SE-ResNet (h-SE-ResNet [44]), CI-BCNN [45] and CSAR [46]. From the results in Tables VII and VIII, we observe that h-SE-ResNet gives the maximum performance among all cases when replaced by our proposed channel attention (88.58% test accuracy on CODEBRIM compared to 89.54% by the proposed iDAAM), however it considers more parameters due to its hierarchical nature (5.14 million compared to 4.89 million by iDAAM), which reflects the superiority of the proposed channel attention mechanism for concrete defect recognition.

Finally, we double the number of channels in each modules (i.e. 64 in first conv, and so on) to observe the change in

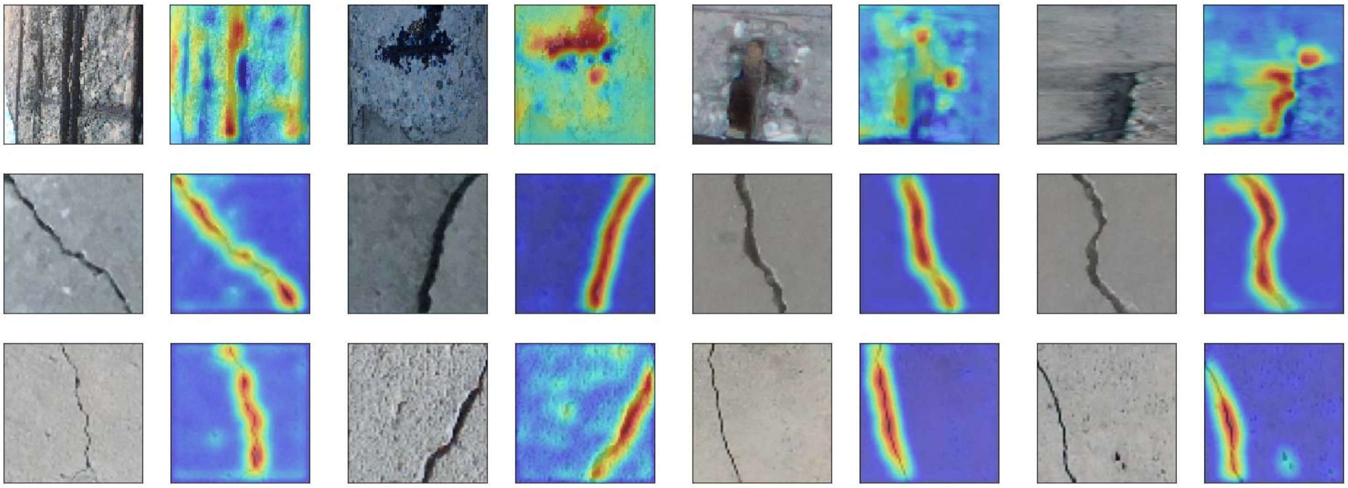

Fig. 4. Attention maps obtained from the proposed iDAAM network for sample images from CODEBRIM, concrete crack image and SDNET-2018 datasets and are given in top, middle and bottom row, respectively. Original images are followed by their respective attention maps, placed side-by-side. Here, red color denotes highest attention, while blue denotes the lowest attention. Top row from left to right: (a) Exposed bars with mild corrosion, efflorescence and spallation, (b) Corroded bar with spallation, (c) advanced spallation with efflorescence and corrosion, (d) Crack surface with spallation. Middle row from left to right: (e)–(h) Images with crack defects from concrete crack image dataset. Bottom row from left to right: (i) Crack surface in a bridge deck, (j) Cracked wall region, (k) Pavement region with crack, (l) Cracked pavement region.

performance when dimension of shrinkage across channels is changed for the global average pooling operation. The experimental results provide similar performance, although considering more number of parameters; thereby showing that the performance of iDAAM does not vary much with the selection of channel dimensions. Similarly, we replace different attention modules with the self-attention block to observe the impact on a self-attention in place of the proposed attention. For all these cases, we obtain reduced performances in Tables VII and VIII, which further demonstrate the importance of the proposed modules.

VI. Analysis and Discussions

A. Analysis Using Attention Maps

Sample images from three datasets are applied on the iDAAM architecture to generate the attention maps by revisiting the global average pooling layer which helps to interpret the decision-making process of the proposed network, as shown in Fig. 4. These attention maps are used to visually illustrate the efficient feature selection mechanism using the proposed modules by highlighting the defective regions, thereby, helping the network to automatically focus on these regions.

Referring to Fig. 4, we can observe that the iDAAM architecture highlights the relevant overlapping defect regions by selecting robust features by the fine-grained dense and concurrent dual attention modules. For example, in Fig. 4 (a), the exposed bars are highlighted with the regions with spallation, efflorescence and corrosion in a sample CODEBRIM image. The crack regions in sample images from concrete crack image dataset have been localized, as shown in Fig. 4 (e)-(h). Crucial regions containing cracks in sample images from bridge deck, wall and pavements of SDNET-2018 dataset are highlighted in Fig. 4 (i)-(l).

TABLE IX

Impact on Recognition Performance by Gradually Stacking Multiple Modules and Sub-Networks on CODEBRIM Dataset

Model Description	Training accuracy	Val. accuracy	Testing accuracy
FGDM + Only spatial attention	98.57	87.68	82.56
FGDM + Only channel attention	98.97	88.59	83.69
FGDM + Only SEM (i.e. channel + spatial)	99.15	90.11	84.27
FGDM + Only CMFA	99.26	89.29	83.56
FGDM + CMFA + Only spatial attention	99.89	91.35	88.93
FGDM + CMFA + Only channel attention	99.92	91.68	89.01
1 FGDM + 1 CDAM	92.58	85.49	75.26
2 FGDM + 2 CDAM	99.94	90.59	88.38
iDAAM	**99.98**	**91.82**	**89.54**

B. Impact of Stacking Multiple Modules for Concrete Defect Recognition

The robust defect feature extraction of iDAAM is attributed to the hierarchy of feature extraction units and concurrent operations of multiple attention architectures. To understand the impact of each modules and the sub-networks, we perform a series of experiments and tabulate the results in Table IX for CODEBRIM dataset. We also generate the attention maps for each case, which are provided in Fig. 5.

We begin the experiments by keeping only the spatial attention part from CDAM with the FGDM and the same has been reciprocated by keeping only channel attention. Figs. 5 (b) and (c) show the attention maps obtained for these cases, respectively. Here, we observe reduction in performance for both cases due to the absence of other attention modules, as found from the attention maps which could not highlight the exposed bars properly. Then we modify iDAAM by keeping both channel and spatial attention with FGDM and removing the CMFA part; resulting in better localization of exposed bars as given in Fig. 5 (d). Then, we replace the CDAM with the CMFA; however, it still couldn't *attend* to spatial and channel

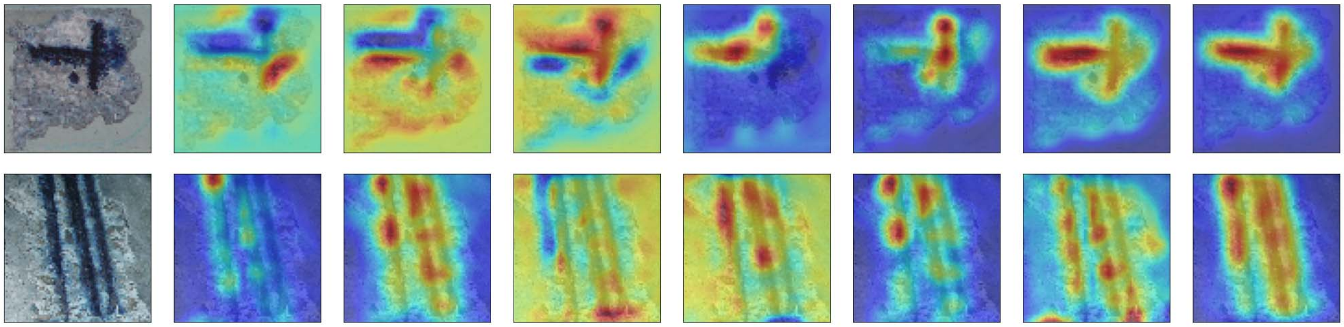

Fig. 5. Attention maps obtained by gradually adding different sub-networks and using well-known CNN architectures on CODEBRIM dataset. First row, from left to right: (a). Defect image, (b). Only spatial attention, (c). Only channel attention, (d). Only SEM, (e) Only CMFA, (f). CMFA + spatial attention, (g). CMFA + channel attention, (h). iDAAM. Second row, from left to right: (i). Defect image, (j). 1 FGDM + 1 CDAM, (k) 2 FGDM + 2 CDAM, (l). Deep CNN with adaptive thresholding, (m). Inception, (n). ResNet-50, (o). DenseNet-121, (p). iDAAM.

locations due to the absence of SEM, reflected in the attention map in Fig. 5 (e).

Then, we modify the CDAM by keeping only one of the spatial and channel attention part in SEM, keeping CMFA unaltered and present the generated attention maps for the same in Fig. 5 (f) and (g), respectively. Here, we observe that the exposed bars are getting more highlighted, whereas the redundant regions are getting darker than the previous ones. Finally, we obtain the attention maps using the full iDAAM architecture in Fig. 5 (h) showing precise localization of defects and successfully suppressing redundant region information. Similarly, we report the results of gradually stacking the blocks and the corresponding attention maps in Table IX and Figs. 5 (j) and (k), respectively. Here, we observe that local features cannot be attended by placing only one FGDM and CDAM, whereas the localization performance substantially increases with stacking.

Furthermore, we show the attention maps generated by training several state-of-the-art models, such as Deep CNN with adaptive thresholding [15], Inception [5], ResNet-50 with transfer learning [6] and DenseNet-121 [7] in Fig. 5 (l)-(o), respectively. Here, we observe that the exposed iron bars are getting moderately noticed using inception [5]; however, the discriminative ability is increased by using deeper networks such as ResNet-50 [6] and DenseNet-121 [7]. Finally, we present the attention map obtained using iDAAM in Fig. 5 (p) which is by the far the best in localization and highlighting defects.

C. Analysis of Single-Class Classification Ability

To understand the network's ability to extract individual class information, experiments are carried out to check the percentage of correctly classifying single class defect out of total five classes. Class specific accuracies are shown in Table X. Experimental results show that the classification accuracy of the exposed bar is higher than other classes, while efflorescence has the lowest tendency to get correctly classified. Table XI shows the performance of model to detect at least one class correctly, then up to 2 classes correctly and so on, up to 4 classes. It is evident from Table XI that up to three classes, the performance is very high and then it degrades when visually-similar small overlapping defect classes (such as efflorescence, spallation and corrosion) are included.

TABLE X
CLASSIFICATION ACCURACY (%) OF SINGLE-CLASS DEFECTS ON CODEBRIM DATASET

Type of defect	Accuracy
Crack	90.15
Spallation	93.56
Efflorescence	88.49
Exposed bars	97.35
Corrosion	89.26

TABLE XI
SINGLE AND MULTI-CLASS CLASSIFICATION ACCURACY (%) ON CODEBRIM DATASET

Number of classes correctly classified	Test accuracy
At least one	100
At least two	99.12
At least three	98.35
At least four	92.78

D. Impact of iDAAM on Retinal Vessel Segmentation

The retinal vessel segmentation is of paramount importance for early diagnosis of several eye-related diseases such as diabetic retinopathy, hypertension, arteriosclerosis, etc. However, several aspects of this problem make it a challenging task, including the presence of small blood vessels, similar appearance in the blood vessel and background and susceptibility towards background lighting and noise. For our case, we have analyzed the impact of the proposed fine-grained dense module and concurrent dual attention module for accurate segmentation of retinal blood vessels.

For this operation, we have considered the U-Net [51] as a backbone network and replaced the second convolution operation at each level with a fine-grained dense module followed by another concurrent dual attention module. We compare the performance of the resulting network to the recent methods in Table XII. For the analysis, we have considered the DRIVE [52] and STARE [53] datasets with image augmentation. Here the proposed network gives better F1-score and AUC than the existing methods in DRIVE dataset. The recent state-of-the-art method [56] uses channel attention in U-Net, whereas the proposed method incorporates multiple attention mechanisms for this purpose, thereby achieving high vessel

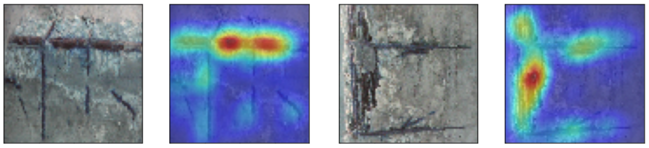

Fig. 6. Examples of failure cases: Attention maps are followed by the original images, placed side by side.

TABLE XII

PERFORMANCE OF iDAAM ARCHITECTURE ON DRIVE [52] AND STARE [53] RETINAL VESSEL SEGMENTATION DATASETS

Method	DRIVE		STARE	
	AUC	**F1-score**	**AUC**	**F1-score**
U-Net [51]	0.9752	0.8174	0.9710	0.7595
DUNet [54]	0.9778	0.8190	0.9758	0.7629
IterNet [55]	0.9813	0.8218	0.9881	0.8146
SA-UNet [56]	0.9864	0.8263	0.9837	0.8175
iDAAM	**0.9875**	**0.8351**	**0.9876**	**0.8458**

segmentation performance. For STARE dataset, it gives better F1-score than the recent works while achieving very similar AUC performance.

E. Discussion on Example Failure Cases

We conclude the analysis of the iDAAM architecture with some examples of failure cases given in Fig. 6 and their probable reasons. To illustrate the failure cases, we have considered two examples, each having all five defect categories being present on the images. The attention maps shown could not localize all defect classes and primarily highlights the most dominant defects of all. This happens where the minute defects get obscured by another dominant defect class. However, we can observe that iDAAM still recognizes the more dominant classes; hence the unhealthy concrete region will not be completely overlooked by iDAAM.

VII. CONCLUSION AND FUTURE WORKS

In this work, we have proposed a solution for automatic multi-target multi-class and single-class classification of defects found in civil concrete infrastructures. Our proposed deep iDAAM architecture is constructed using FGDM and CDAM, which are interleaved to extract robust salient discriminative features from multiple scales to improve the classification performance with less parameters. Extensive experimental results and ablation studies show that the proposed iDAAM architecture outperforms many state-of-the-art methods on three large datasets: CODEBRIM, Concrete crack image dataset and SDNET-2018. In particular, for the difficult multi-target multi-class classification problem, it achieves multi-target accuracy in CODEBRIM dataset as high as 89.54%, compared to 84.29% by the current state-of-the-art method. The effectiveness of the proposed iDAAM architecture can be effective for classification of other types of defects within and outside concrete and steel structures. Also, the iDAAM solution can be utilized in conjunction with unmanned aerial vehicles (UAVs) for fast and accurate damage detection, health prediction and monitoring of massive concrete structures.

REFERENCES

[1] (2018). *Morandi Bridge Collapse*. [Online]. Available: https://en.wikipedia.org/wiki/Ponte_Morandi#Collapse

[2] B. Phares. (2001). *Proceedings: Annual Meeting of TRB Subcommittee A2C05 (1)*. Report No FHWARD01-020 and FHWA-RD-01-0212001. [Online]. Available: https://www.fhwa.dot.gov/publications/research/nde/pdfs/01020a.pdf

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Las Vegas, NV, USA, Dec. 2012, pp. 1106–1114.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–14.

[5] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Nov. 2016, pp. 770–778.

[7] G. Huang, Z. Liu, L. Maaten, and K. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269.

[8] V. Andrearczyk and P. F. Whelan, "Using filter banks in convolutional neural networks for texture classification," *Pattern Recognit. Lett.*, vol. 84, pp. 63–69, Dec. 2016.

[9] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, York, U.K., Sep. 2016, p. 87.1–87.12.

[10] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack classification using random structured forest," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3434–3445, Dec. 2016.

[11] L. Yang, B. Li, W. Li, Z. Liu, G. Yang, and J. Xiao, "Deep concrete inspection using unmanned aerial vehicle towards CSSC dataset," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Vancouver, BC, Canada, Sep. 2017, pp. 24–28.

[12] S. Dorafshan, R. J. Thomas, and M. Maguire, "SDNET2018: An annotated image dataset for non-contact concrete crack detection using deep convolutional neural networks," *Data Brief*, vol. 21, pp. 1664–1668, Dec. 2018.

[13] B. Kim and S. Cho, "Automated vision-based classification of cracks on concrete surfaces using a deep learning technique," *Sensors*, vol. 18, no. 10, pp. 34–52, Oct. 2018.

[14] N. Wang, Q. Zhao, S. Li, and X. Zhao, "Damage classification for masonry historic structures using convolutional neural networks based on still images," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, pp. 1073–1089, Aug. 2018.

[15] R. Fan *et al.*, "Road crack detection using deep convolutional neural network and adaptive thresholding," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Paris, France, Jun. 2019, pp. 474–479.

[16] M. Mundt, S. Majumder, S. Murali, P. Panetsos, and V. Ramesh, "Meta-learning convolutional neural architectures for multi-target concrete defect classification with the COncrete DEfect BRidge IMage dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 11196–11205.

[17] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 6000–6010.

[18] S. H. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[19] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458.

[20] X. Yang, D. Lo, X. Xia, Y. Zhang, and J. Sun, "Deep learning for just-in-time defect prediction," in *Proc. IEEE Int. Conf. Softw. Qual., Rel. Secur.*, Vancouver, BC, Canada, Aug. 2015, pp. 17–26.

[21] S. Wang, T. Liu, and L. Tan, "Automatically learning semantic features for defect prediction," in *Proc. 38th Int. Conf. Softw. Eng.*, Austin, TX, USA, May 2016, pp. 297–308.

[22] J. Park, J. Lee, S. Woo, and I. Kweon, "BAM: Bottleneck attention module," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Newcastle, U.K., Sep. 2018, pp. 92.1–92.14.

[23] J. Park, J. Lee, S. Woo, and I. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 3–19.

[24] Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao, and H. Lu, "Attention CoupleNet: Fully convolutional attention coupling network for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 113–126, Jan. 2019.

[25] S. Zhou *et al.*, "Hierarchical U-shape attention network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 8417–8428, Jul. 2020.

[26] R. Ye, C.-S. Pan, M. Chang, and Q. Yu, "Intelligent defect classification system based on deep learning," *Adv. Mech. Eng.*, vol. 10, no. 3, Mar. 2018, Art. no. 168781401876668.

[27] D. Purwanto, R. R. A. Pramono, Y.-T. Chen, and W.-H. Fang, "Three-stream network with bidirectional self-attention for action recognition in extreme low resolution videos," *IEEE Signal Process. Lett.*, vol. 26, no. 8, pp. 1187–1191, Aug. 2019.

[28] H. Yang, C. Yuan, L. Zhang, Y. Sun, W. Hu, and S. J. Maybank, "STA-CNN: Convolutional spatial-temporal attention learning for action recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 5783–5793, Apr. 2020.

[29] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1347–1360, Mar. 2018.

[30] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6298–6306.

[31] J. Ji, C. Xu, X. Zhang, B. Wang, and X. Song, "Spatio-temporal memory attention for image captioning," *IEEE Trans. Image Process.*, vol. 29, pp. 7615–7628, Jun. 2020.

[32] F. Huang, X. Zhang, Z. Zhao, and Z. Li, "Bi-directional spatial-semantic attention networks for image-text matching," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2008–2020, Apr. 2019.

[33] H. Guo, K. Zhu, M. Tang, and J. Wang, "Two-Level attention network with multi-grain ranking loss for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4328–4338, Sep. 2019.

[34] G. Chen, J. Lu, M. Yang, and J. Zhou, "Learning recurrent 3D attention for video-based person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 6963–6976, May 2020.

[35] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[36] Z. Tan, Y. Yang, J. Wan, H. Hang, G. Guo, and S. Z. Li, "Attention-based pedestrian attribute analysis," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6126–6140, Dec. 2019.

[37] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, "Road crack detection using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 3708–3712.

[38] C. V. Dung and L. D. Anh, "Autonomous concrete crack detection using deep fully convolutional neural network," *Autom. Construct.*, vol. 99, pp. 52–58, Mar. 2019.

[39] M. Słoński, "A comparison of deep convolutional neural networks for image-based detection of concrete surface cracks," *Comput. Assist. Methods Eng. Sci.*, vol. 26, no. 2, pp. 105–112, Feb. 2019.

[40] H. Pham, M. Y. Guan, B. Zoph, Q. V. Le, and J. Dean, "Efficient neural architecture search via parameter sharing," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, Jul. 2018, pp. 4095–4104.

[41] B. Baker, O. Gupta, N. Naik, and R. Raskar, "Designing neural network architectures using reinforcement learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017, pp. 1–18.

[42] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[43] H. Zhang *et al.*, "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*. [Online]. Available: http://arxiv.org/abs/2004.08955

[44] C. Li, Z. Chen, Q. M. J. Wu, and C. Liu, "Deep saliency detection via channel-wise hierarchical feature responses," *Neurocomputing*, vol. 322, pp. 80–92, Dec. 2018.

[45] Z. Wang, J. Lu, C. Tao, J. Zhou, and Q. Tian, "Learning channel-wise interactions for binary convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 568–577.

[46] Y. Hu, J. Li, Y. Huang, and X. Gao, "Channel-wise and spatial feature modulation network for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3911–3927, Nov. 2020.

[47] G. Bhattacharya, B. Mandal, and N. B. Puhan, "Multi-deformation aware attention learning for concrete structural defect classification," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Sep. 30, 2020, doi: 10.1109/TCSVT.2020.3028008.

[48] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 1, 2020, doi: 10.1109/TPAMI.2020.2983686.

[49] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 483–499.

[50] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 936–944.

[51] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany, Oct. 2015, pp. 234–241.

[52] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, and B. van Ginneken, "Ridge based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[53] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans. Med. Imag.*, vol. 19, no. 3, pp. 203–210, Mar. 2000.

[54] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "DUNet: A deformable network for retinal vessel segmentation," *Knowl.-Based Syst.*, vol. 178, pp. 149–162, Aug. 2019.

[55] L. Li, M. Verma, Y. Nakashima, H. Nagahara, and R. Kawasaki, "IterNet: Retinal image segmentation utilizing structural redundancy in vessel networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Snowmass, CO, USA, Mar. 2020, pp. 3645–3654.

[56] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, "SA-UNet: Spatial attention U-Net for retinal vessel segmentation," Apr. 2020, *arXiv:2004.03696*. [Online]. Available: http://arxiv.org/abs/2004.03696

[57] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang, "Visual attention consistency under image transforms for multi-label image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 729–739.

[58] G. Bhattacharya, B. Mandal, and N. B. Puhan. (2021). *Interleaved Artifacts Aware Attention Mechanism (Source Code)*. [Online]. Available: https://github.com/NBPuhan/iDAAM

Gaurab Bhattacharya received the B.Tech. degree in electronics and communication engineering (ECE) from the Institute of Engineering and Management (IEM), Kolkata, India, in 2018, and the M.Tech. degree in ECE from the Indian Institute of Technology (IIT) Bhubaneswar, India, in 2020. His research interests include image processing, computer vision, machine learning, and biometrics.

Bappaditya Mandal received the B.Tech. degree in electrical engineering from the Indian Institute of Technology (IIT) Roorkee, India, in 2003, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2008. He is currently a Lecturer in computer science with the School of Computing and Mathematics, Keele University, U.K. He has worked as a Scientist with the Visual Computing Department, Institute for Infocomm Research, A*STAR, Singapore. His research interests include subspace learning, feature extraction and evaluation, computer vision, image and signal analysis, and machine learning. He is a member of the IEEE Signal Processing Society and the British Computer Society.

Niladri B. Puhan (Member, IEEE) received the B.E. degree in electrical engineering from UCE Burla, the M.E. degree in signal processing from the Indian Institute of Science, Bengaluru, India, and the Ph.D. degree from Nanyang Technological University, Singapore. He is currently an Assistant Professor with the School of Electrical Sciences, IIT Bhubaneswar. His fields of research interests include signal and image processing, computer vision, biometrics, machine learning, and medical image analysis.