Signal, Image and Video Processing (2024) 18:485–494
https://doi.org/10.1007/s11760-023-02745-2

ORIGINAL PAPER

A novel multi-scale cross-patch attention with dilated convolution (MCPAD-UNET) for metallic surface defect detection

Ali Furkan Kamanli[1]

Received: 27 July 2023 / Revised: 3 August 2023 / Accepted: 13 August 2023 / Published online: 27 September 2023

Abstract
Surface defect detection in industrial processes is crucial for ensuring product quality and reducing material waste. Automated defect identification using deep learning techniques has become a vital aspect of the automated surface defect detection field. However, achieving accurate and automatic defect segmentation remains a significant challenge, especially for fine precision segmentation required in high-quality products. The traditional approaches for defect segmentation have several limitations, such as difficulty in preserving fine details and contextual information, leading to poor segmentation performance. To overcome these limitations, new segmentation algorithms that can preserve fine precision and contextual information need to be evaluated. Therefore, there is a need for novel segmentation algorithms that can accurately identify and segment defects in industrial processes, incorporating multi-scale contextual information, preserving fine details, and handling complex and subtle defects. In this paper, we propose a novel approach for steel defect segmentation called multi-scale cross-patch attention with dilated convolution (MCPAD-UNet). This approach employs a subsampled module that achieves the same dimensionality reduction as max-pooling while preserving the fine precision of the features. Additionally, MCPAD-UNet utilizes a cross-patch attention module with dilated convolution, simultaneously collecting channel–spatial data and integrating relevant multi-scale features to reduce the semantic gap and enhance detailed information. To prevent overfitting, we apply dropout after each hybrid dilated convolution block. Extensive testing on the public Severstal: Steel Defect Detection dataset demonstrates the effectiveness of our approach, achieving Dice scores of 95.3%, outperforming the competition's overall score by 5.2%. Our proposed method has the potential to significantly improve defect detection in industrial processes, thereby reducing material waste and improving product quality.

Keywords Deep learning · Segmentation · Attention · Surface defect

1 Introduction

Machine vision algorithms for surface defect analysis, especially for non-metallic surfaces, have garnered significant attention in recent years. Two main types of approaches have been studied: conventional computer vision and pattern recognition methods employing superficial learning or unique features [1]. Conventional image analysis techniques for detecting and segmenting abnormalities rely on basic local abnormality characteristics, categorized as structural, threshold, spectral, or model-based techniques [2]. Structural techniques include edge identification, skeletonization, template matching, and morphological procedures [3–5].

Several CNN-based techniques for surface defect detection have been developed, each with its own advantages and disadvantages [6]. In the literature, various approaches, such as multi-scale pyramidal pooling networks, adaptable multi-layered deep feature extraction frameworks, and automated quality visual assessment procedures using conventional CNNs with sliding windows, have been proposed [7]. While many deep convolutional neural network (DCNN)-based detection systems and methods for structural damage detection have been introduced, some of these techniques suffer from using bounding boxes, leading to imprecise defect boundary localization. To address these limitations, researchers have proposed deep learning-based approaches that employ pre-trained networks to classify defect image

✉ Ali Furkan Kamanli
fkamanli@sakarya.edu.tr

[1] Faculty of Technology, Electrical and Electronics Engineering, Sakarya University of Applied Sciences, Sakarya, Turkey

patches and subsequently use segmentation methods for pixel-wise defect predictions [8, 9].

One commonly used segmentation method is the U-Net, a novel U-shaped architecture with a symmetric topology [10]. However, the U-Net faces challenges in handling the semantic gap between coarse-grained features, making it difficult to separate tiny defects. Additionally, the U-Net's subsampled operation reduces segmentation accuracy by losing information [11]. To overcome these issues, modifications to the U-Net architecture have been proposed, including the Channel Attention SE block, SA-UNet, attention gates, Channel-UNet, and multiscale feature fusion. Recently, UNet++, composed of UNets of various depths and utilizing dense connections, has shown improved segmentation performance [12]. Nevertheless, it still struggles with retrieving coarse-grained characteristics while considering the semantic gap, leading to difficulties in separating small defects. Addressing these challenges, researchers have introduced various U-Net modifications [13], such as the Channel Attention SE block, SA-UNet, bottleneck feature-driven U-Net, attention gate, Channel-UNet, multiscale feature fusion, Res2Net module, MS-UNet, and ResDUNet [14]. To tackle the issue of contextual fusion data scarcity, the CPAD-Net: Contextual Parallel Attention and Dilated Network was proposed for liver tumor segmentation, demonstrating accurate segmentation with high precision [15]. Similarly, the CPAM: Cross-Patch Attention Module employs a multi-scale attention mechanism and cross-patch attention mechanism for detecting tile defects with object detection algorithms [16], showing potential for segmenting metal defects using hybrid mechanisms.

In this research, we propose a new technique, MCPAD-UNet, for accurately segmenting steel defects, incorporating multi-scale contextual cross-patch attention with dilated convolution.

We evaluated our method on the Severstal: Steel Defect Detection dataset, part of a Kaggle competition, achieving Dice scores of 95.3%, outperforming the competition's overall score by 5.2%. Our proposed technique demonstrates computational efficiency and robustness, making it suitable for real-time and accurate detection of steel defects in industrial applications. Moreover, our approach can be extended to various industrial settings that require precise and automatic defect segmentation, not limited to steel defects.

2 Materials and methods

In this section, we explain a novel multi-scale contextual cross-patch attention with dilated convolution (MCPAD-UNET) approach. We provide an overview of several relevant studies in Supplementary Table 1, summarizing the general literature and key findings briefly. Various key factors need to be taken into consideration.

The studies have demonstrated that deep learning-based algorithms, such as CNNs, Mask RCNN, and UNET, can accurately detect surface defects in metallic parts. The model's performance and accuracy can be enhanced by utilizing methods such as transfer learning, GAN-based data augmentation, attention mechanisms, ensemble learning, and feature fusion.

Frequently utilized in image segmentation, the UNet architecture employs a contracting path to capture context and a symmetric expanding path to achieve precise localization. CPAD-Net utilizes dilated convolutions to broaden the network's receptive area [15] and collect contextual data at various sizes. In contrast, for complex texture tile block defect identification, CPAM improves the characteristics pertinent to defects by collecting contextual information through its spatial and channel attention techniques [16]. The incorporation of these modules into the UNet architecture can improve the model's ability to detect complex texture defects in images and capture long-range correlations. It is crucial to note that the success of our hybrid strategy for detecting metallic surface defects depends on the precise image segmentation objective we decide to utilize. The computational cost of integrating these extra modules into the UNet design must also be considered.

The integration of CPAM and CPAD-Net with the UNet model has the potential to enhance image segmentation tasks by capturing more contextual information and detecting complex texture defects. Ablation experiments are conducted to determine the optimal architecture for a specific image segmentation task, such as metallic surface defect detection (using the Severstal: Metallic Surface Defect Detection Dataset [27]).

2.1 Surface defect dataset

The Severstal Metallic Surface Defect Detection dataset is a computer vision dataset containing images of metallic surfaces with various types of defects [27]. Severstal, one of the top mining firms for steel and related products, created the dataset to enhance quality control procedures and lower production costs.

The dataset comprises nearly 5000 high-quality images of metallic surfaces, each with a resolution of 1600×256 pixels. It is divided into two sections: a training set with 4000 images and a test set with 1000 images. Each image in the dataset is labeled at the pixel level to indicate imperfections such as cracks, scratches, and other surface abnormalities.

The Severstal Metallic Surface Defect Detection dataset has become the foundation for several research articles and machine learning contests aimed at creating more precise and effective models for surface defect detection in industrial settings. Supplementary Figure 1 demonstrates examples of images from the dataset used in these studies.

2.2 MCPAD-UNet

Convolutional neural networks (CNNs) of the U-Net variety are frequently employed for image segmentation tasks. U-Net was initially created to segment medical images, but it has since found applications in various domains, including industrial image segmentation [23–25].

According to the literature, "CPAM: Cross-Patch Attention Module for Complex Texture Tile Block Defect Detection" provides an attention technique to improve the detection of complex texture tile block defects [16]. The authors describe the Cross-Patch Attention Module (CPAM), which combines channel attention with spatial attention. To develop a novel MCPAD-UNet model for multiclass surface defect detection, the authors integrate CPAM [16] and CPAD-Net [15] modules and make some optimizations.

A novel deep-learning architecture for segmenting metallic defects is proposed by MCPAD-Net. The suggested architecture comprises a contextual parallel attention module and a dilated convolutional network. The study evaluates MCPAD-UNet's performance and compares it to cutting-edge techniques using a public Severstal Metallic Surface Segmentation dataset.

The MCPAD-UNet architecture is visually demonstrated in Fig. 1.

The U-Net utilizes 2×2 max pooling for subsampled data with a stride of 2. However, this approach may hinder the localization of tiny surface defects due to reduced image quality and loss of small-volume defect information caused by max pooling. In contrast, our study adopts a practical subsampled module [15], as illustrated in Supplementary Fig. 3, which effectively reduces the feature map size by half without losing information. This reduction is achieved by setting both the convolution kernel size and stride to 2×2. To accelerate training and convergence, Batch Normalization (BN) is applied after convolution. Additionally, the ReLU function is employed to introduce non-linearity and prevent vanishing gradients, as it is commonly known to enhance the transformation of networks.

2.3 Hyperparameter optimization

Deep learning model training must include hyperparameter adjustment, and a dataset's performance may be assessed using a variety of methods. In this work, we evaluated the performance of the dataset using AdamW, Random Sampler, Focal Loss, Noam Scheduler, and Fine Tuning [28, 29].

In binary classification tasks, the binary cross-entropy (BCE) loss function is frequently employed, which is determined by taking the negative logarithm of the projected probability of the positive class [30].

To address this issue, the Focal Loss function was suggested [31].

The BCE + Focal loss function combines the BCE and Focal loss functions, using the former to calculate the loss for low-level features and the latter for high-level features.

Mathematically, the BCE loss + Focal loss can be stated as follows: Let p be the expected probability of the positive class, and let y represent the ground truth label, which can take the values 0 or 1 (i.e., class 1). The loss due to binary cross-entropy (BCE) is provided by:

$$\mathrm{BCE}(y,\ p) = -y * \log(p) - (1 - y) * \log(1 - p) \tag{1}$$

The Focal loss, which de-weights the loss allocated to samples with proper classification, was created to solve the problem of class imbalance. The Focal Loss is stated as follows:

$$\mathrm{FL}(y,\ p) = -\alpha * (1 - p)\hat{}\gamma * y * \log(p) - (1 - \alpha) * p\hat{}\gamma * (1 - y) * \log(1 - p) \tag{2}$$

Here, the focusing parameter enhances the contribution of high-level characteristics while down-weighting the contribution of low-level features. The balancing parameter governs the trade-off between positive and negative samples.

As a result, the total BCE loss + Focal loss may be expressed as:

$$\mathrm{BCE} + \mathrm{FL}(y,\ p) = \lambda * \mathrm{BCE}(y,\ p) + (1 - \lambda) * \mathrm{FL}(y,\ p) \tag{3}$$

Here, λ is a hyperparameter that regulates the compromise between BCE and FL. The hyperparameter optimization process can be utilized to determine the optimal value of λ.

2.4 Performance metrics

The effectiveness of semantic segmentation models like UNET is often evaluated using the F1 score and the Intersection over Union (IoU) metric [25]. The F1 score calculates the harmonic mean of accuracy and recall using the projected and ground-truth segmentation masks [26]. To elaborate, recall evaluates the proportion of real positive pixels to the total number of projected positive pixels, while precision evaluates the proportion of real positive pixels to the total number of real positive pixels. A higher F1 score indicates better segmentation performance.

The IoU, also known as the Jaccard index [27], measures the overlap between the expected and actual segmentation masks.

For binary classification models like the UNET used for picture segmentation, the F1 and IoU scores are popular statistics, and they are determined by taking the harmonic

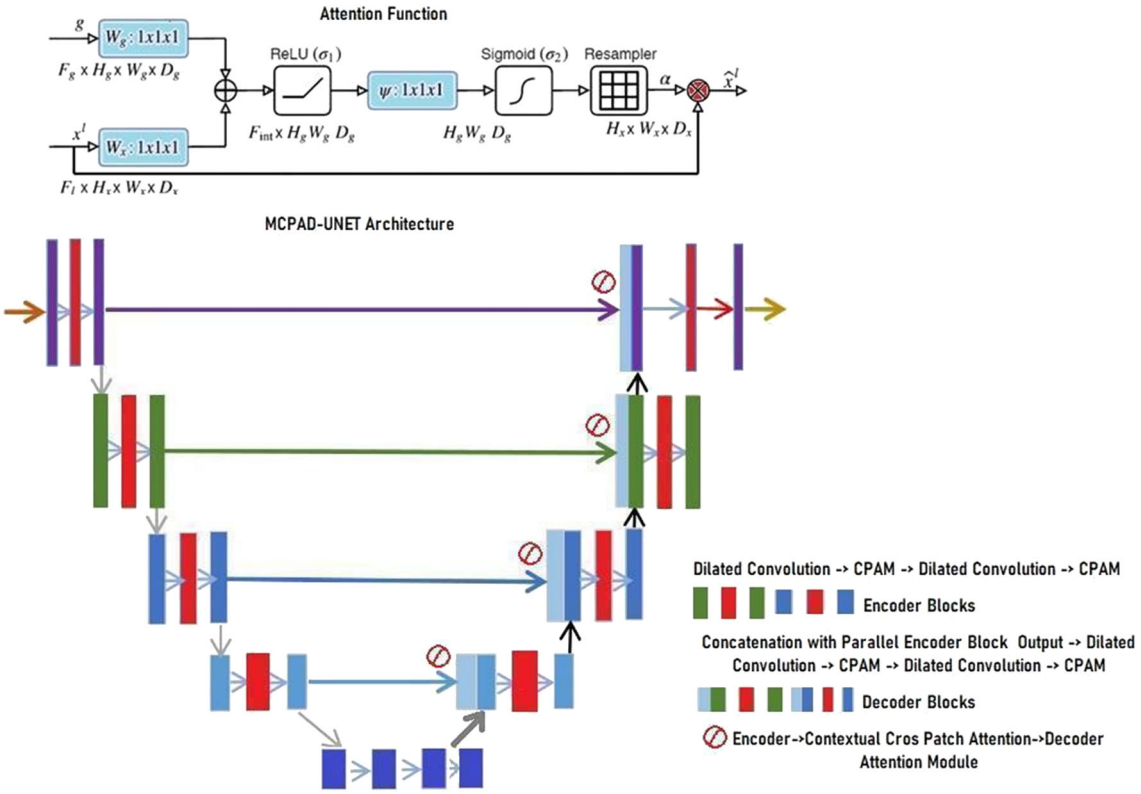

Fig. 1 A novel MCPAD-UNET architecture

mean of recall and precision.

$$\text{F1 score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}),$$
$$\text{IOU} = \text{intersection/union}$$

In the context of image segmentation, recall represents the percentage of correctly predicted positive instances among all positive cases.

2.5 Pre-processing and data augmentation

Augmentations are techniques used to increase the size and diversity of a machine learning model's training dataset. By applying various modifications to the input data, augmentations create additional training instances from the existing data. This process aids in enhancing the model's generalization to new examples by exposing it to a wider range of input data during training. It is essential to select appropriate augmentations based on the likelihood of their occurrence. Augmentations such as horizontal flip, vertical flip, rotation, translation, and scaling may occur depending on the camera angle, while GaussNoise, ElasticTransform, and RandomBrightnessContrast might be present due to camera situations and environmental conditions. In this study, augmentation techniques were applied accordingly [32]. In surface defect detection tasks, data with such variations are commonly encountered [33], making augmentations highly valuable, as depicted in Supplementary Fig. 4.

3 Results and discussion

Severstal is a dataset released as part of a Kaggle competition in 2019, consisting of images of steel sheets. The objective of the competition was to develop artificial intelligence models capable of identifying defects in steel sheets. The dataset provides labeled images, with annotations specifying the type and location of any defects present. Firstly, exploratory data analysis was conducted and is presented in Supplementary Fig. 1 and Fig. 1.

Prior to feeding the dataset into deep learning models, pre-processing is performed to eliminate noise. The dataset, compiled from various sources, may contain unwanted distortions, making the analysis impractical. Additionally, the data comprise properties with varying scales. Rescaling the characteristics is beneficial for deep learning algorithms, as it aids in a faster learning process. Consequently, to ensure uniform sizes and resolutions across all images, the images must be appropriately rescaled. For image classification, rescaling the input images to an appropriate size is crucial [7], as images that are too small might lead to overlapping, while excessively large images prolong the training process.

OpenCV (Open Source Computer Vision) is used for the preprocessing stage.

For multi-class analysis of surface defect images, MCPAD-UNet is employed. To utilize the BCE + Focal loss function with MCPAD-UNet, the final layer of the network needs to be modified to output probabilities for each class, followed by the application of the loss function to these probabilities.

To compute the overall loss for the entire image, you can simply average the BCE + Focal loss values for all pixels in the image. Backpropagate the error and update the weights. Once you have calculated the overall loss for the image, you can backpropagate the error through the network and update the weights.

MCPAD-UNet architecture: Input Image → MCPAD-UNet → Softmax Activation → Sigmoid Activation → Probability Maps → BCE + Focal Loss → Average Loss → Backpropagation → Update Weights.

In Supplementary Fig. 3 and the architecture diagram, the MCPAD-UNet architecture processes the input image, producing a collection of probability maps for each class.

The Severstal dataset, comprising labeled images of steel sheets annotated with the type and location of defects, was used in this study. The exploratory data analysis of the dataset is presented in Supplementary Figs. 5 and 6.

Several possible outcomes can be derived from the Severstal dataset:

- A large number of samples belong to class 3, resulting in a highly imbalanced dataset where almost 73% of all defects are from class 3.
- Class 4 defects account for 11.3% of all defects, but they cover nearly 17% of the total defect area. This suggests that class 4 defects often have larger sizes.
- Classes 1 and 2 exhibit relatively smaller defects, with 12.6% and 3.48% of all defects falling into classes 1 and 2, respectively. However, they only account for 2.39% and 0.51% of the total number of pixels in the defect mask, respectively.
- Deep learning models encounter challenges in identifying class 1 and class 2 data due to their small sizes.
- The box plot further confirms the previous observation that class 4 defects are generally larger in size compared to class 3, as well as class 1 and 2.
- Defect class 3 shows many outliers. Despite class 4 defects being generally larger in size, the outlier values in class 3 can be much larger than those in class 4.

Class 1 is often tiny in size, composed of several components, and has the largest percentage of segments with more than five segments, as well as the smallest total defect area. The mask size and threshold correlation are shown in Supplementary Fig. 7.

The possibility of many fault classes was explored using a method known as FP (Frequent Pattern) growth to determine the kinds of defects that frequently occur together:

- The frequency chart above shows that a picture with a single defect of class 3, 1, or 4 is the most common case.
- Classes 3 and 4 appear more frequently together than class 2 alone. This observation is intriguing, especially considering that classes 3 and 1 make up the most common samples in the dataset.
- To increase the number of examples for class 2, more augmentation is needed.

The performance of model training was enhanced using these augmentation strategies.

3.1 Ablation study for surface defect detection in steel sheets using MCPAD-UNet

In this ablation study, we aim to analyze the impact of different components and modules in our proposed deep learning model, MCPAD-UNet, for surface defect detection in steel sheets. We will gradually enable or disable specific components and measure the performance on the Severstal: Steel Defect Detection dataset. The components we will focus on are as follows:

1. Hybrid Dilation Convolution (HDC):
   - The HDC block is responsible for enhancing the encoding stage of the MCPAD-UNet by using dilated convolutions at different scales for feature extraction.
   - We will compare the performance of the model with and without the HDC block to evaluate its effectiveness.
2. Double Dilated Convolution (DDC):
   - The DDC block is used in the decoding stage of the MCPAD-UNet to increase the receptive field and extract multi-scale features.
   - We will analyze the impact of DDC on the segmentation performance by enabling or disabling it in the network.
3. Channel and Spatial Attention (CPAM):
   - The CPAM module is integrated into the MCPAD-UNet to calibrate the fused feature maps at skip connections, focusing on dominant channel and spatial information.
   - We will evaluate the contribution of CPAM by comparing the results with and without this attention mechanism.
4. Dropout Regularization:

- Dropout layers are added after each DDC block to prevent overfitting during training.
- We will assess the effect of dropout by varying the dropout rate and observing its impact on the overall segmentation performance.

The ablation study will provide valuable insights into the contributions of each component and module in the MCPAD-UNet architecture for surface defect detection in steel sheets. It will help us understand which components are crucial for achieving superior segmentation performance and robustness. The results will guide us in selecting the most effective configuration of the MCPAD-UNet and shed light on potential improvements for future research in industrial defect detection applications.

3.2 Training results comparison

The proposed deep learning network in this study is created using the PyTorch framework and trained on an Intel I7-12700k CPU and an Nvidia GeForce RTX 3070 graphics card with 12 GB of RAM. The network's hyperparameters are optimized for maximum performance based on prior knowledge. The setup involves training the network with the Adam optimizer over 100 epochs using a mini-batch size of 16 and an initial learning rate of 1e−4. Training is terminated if the loss value does not decrease after 30 epochs. The network utilizes a hybrid dilation convolution (HDC) and double dilated convolution (DDC) with dilation rates of 1, 3, 5 and 7 and a convolution kernel size of 3×3.

The experimental results demonstrate that the best segmentation performance is achieved when HDC is used during the encoding stage and DDC during the decoding stage, resulting in a Dice score of 69.81%, a VOE of 33.3%, and an RVD of 3.9%. Moreover, the segmentation performance of the encoder is significantly enhanced by employing dilated convolutions with four different scales for feature extraction, resulting in a 2.49% increase in the Dice score, which is considered a notable improvement.

A Dropout layer (also known as U-Net + Dropout) is added after the double dilated convolution (DDC) block to improve the network's robustness. The impact of the dropout rate (p) on the network's performance is evaluated, and the best p is found to be 0.37 (Dice: 88.28%), producing the best segmentation performance.

Ablation tests are conducted on the proposed network's backbone, which includes subsampled (Sb), dilated convolution (Dc), and channel attention (CPAM) blocks. The results in Table 1 show that as modules are gradually added to the backbone, the segmentation performance significantly improves. The addition of the CPAM block leads to the largest performance improvement, increasing the Dice score by 7.78% and providing the best segmentation performance for the MCPAD-Net configuration. The CPAM block is found to be essential to the proposed network, improving segmentation performance by about 9% compared to the backbone.

The segmentation performance improves as additional modules are added, but it also results in an increase in network depth, the number of parameters, and computational complexity (measured in GFLOPs). Specifically, applying the Sb, Dc, and CPAM blocks to the backbone enhances segmentation performance by 3.58%, 8%, and 7.58%, respectively (as shown in Table 1). The use of a large number of dilated convolutions by the Dc block, which improves the network's perceptual abilities, is found to be the most effective way to increase segmentation performance.

The proposed network combines dilated convolution, UNet main portion, and CPAM in several different backbones, including UNet, UNet + Sb, UNet + Dc, UNet + CPAM, CPAD-Net, and MCPAD-UNet. Comparing these networks to the network outlined in [15], the proposed network achieves a Dice score of 91.78%, a VOE of 31.9%, and an RVD of only 5.89%. Our findings indicate that MCPAD-UNet outperforms all rivals, including competition outcomes [27] (Severstal). The CPAM block in MCPAD-UNet enhances the potential for gathering channel and spatial information, in addition to paying attention to contextual information between slices. Overall, the experimental findings demonstrate that all of the suggested modules significantly enhance the network's segmentation performance, as depicted in Supplementary Fig. 8.

In summary, we proposed a novel hybrid attention mechanism U-Net architecture, called MCPAD-UNet, for steel defect segmentation. The proposed network integrates multi-scale contextual cross-patch attention with dilated convolution to enhance detailed information while reducing the semantic gap. We trained our proposed network on the Severstal: Steel Defect Detection dataset, and achieved a Dice score of 95.3%, outperforming the competition's overall score by 5.2%. The segmentation examples are shown in Fig. 2, and the heatmap of the MCPAD-UNet attention is shown in Supplementary Fig. 10.

The proposed approach has significant potential to revolutionize automated defect segmentation in industrial settings, where it can reduce material waste and enhance product quality. Future research will concentrate on increasing the suggested method's computational effectiveness and assessing overall performance on various datasets and in other commercial applications.

4 Discussion

MCPAD-UNet is a novel deep learning architecture proposed for surface defect detection in industrial steel sheets. It builds upon the classic U-Net architecture by introducing several

Table 1 Comparison of several backbone architectures for MCPAD-UNET on Severstal

BackBone	PA (%)	MIoU (%)	FWIoU (%)	DC (%)
ResNet52	90.14	81.14	84.23	90.51
ResNet102	94.17	84.16	88.74	92.36
ResNet152	97.69	90.62	96.52	95.35
VGG16	89.78	80.61	85.74	86.52
DenseNet169	93.56	89.34	92.54	90.50
Xception	92.25	82.12	85.36	83.20

PA, Pixel accuracy; MIoU, Mean IoU; FWIoU, Frequency weighted IoU, DC, dice coefficient

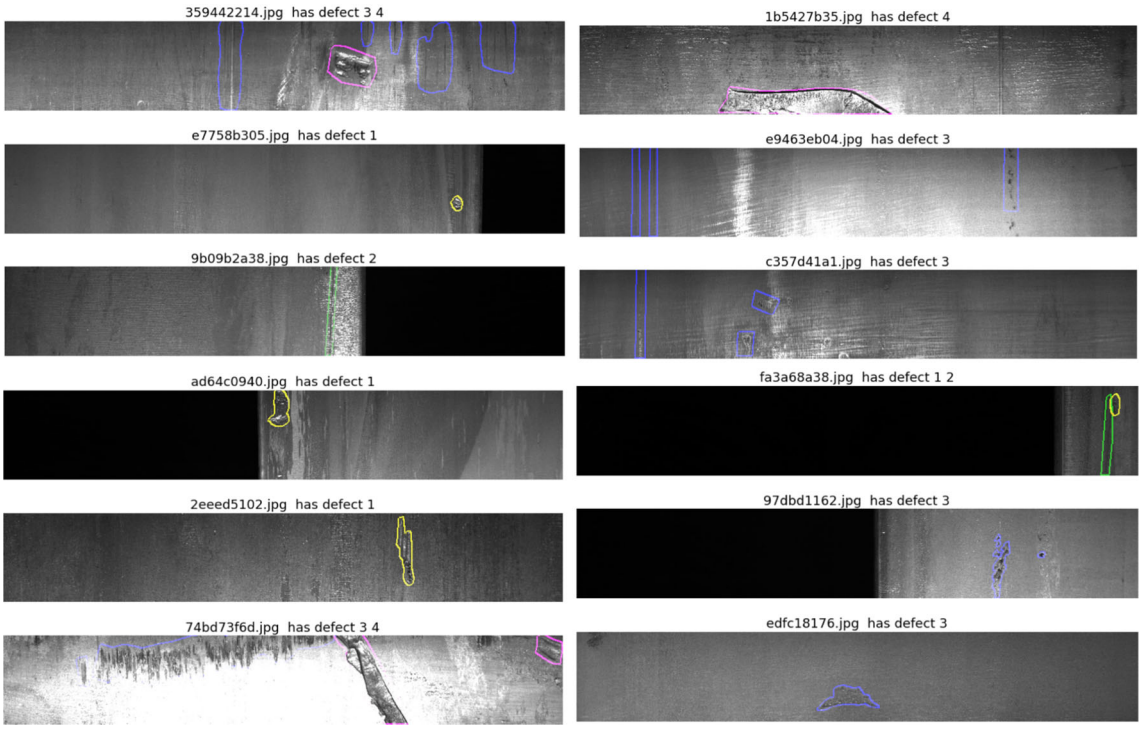

Fig. 2 Detection examples with multiclass defects with MCPAD-UNet

innovative components, including hybrid dilation convolution (HDC), double dilated convolution (DDC), and channel and spatial attention (CPAM) modules.

- MCPAD-UNet versus U-Net: U-Net is a widely used encoder–decoder architecture for image segmentation, particularly in biomedical applications. While U-Net has shown good performance in various segmentation tasks, it may struggle to capture fine-grained details and context in complex images like industrial steel sheets. In contrast, MCPAD-UNet addresses these limitations by incorporating HDC and DDC blocks, enabling the extraction of multi-scale features and the effective handling of larger input sizes.

The introduction of the CPAM module further enhances the performance of MCPAD-UNet by providing attention calibration and feature refinement. This attention mechanism allows the model to focus on relevant information, leading to more accurate defect segmentation. As a result, MCPAD-UNet is likely to outperform the traditional U-Net in surface defect detection tasks, especially when dealing with intricate defects and varying defect sizes.

- MCPAD-UNet versus SegNet: SegNet is another encoder–decoder architecture for semantic segmentation that uses pooling indices for upsampling, making it computationally efficient. However, this pooling-based

upsampling may not be ideal for tasks requiring precise localization and capturing fine details. In contrast, MCPAD-UNet utilizes dilated convolutions and attention mechanisms, making it better suited for handling complex images with large input sizes.

- MCPAD-UNet versus DeepLabv3: DeepLabv3 is a state-of-the-art deep learning architecture for semantic segmentation, known for its ability to capture fine details using atrous (dilated) convolutions. While DeepLabv3 has demonstrated impressive performance in various image segmentation challenges, it can be computationally expensive, especially with large input sizes.

MCPAD-UNet shares some similarities with DeepLabv3 in leveraging dilated convolutions for multi-scale feature extraction.

- MCPAD-UNet versus Mixed Supervision: Mixed supervision is an approach that combines both fully supervised and weakly supervised learning for training deep learning models. In the context of surface defect detection, mixed supervision uses pixel-level annotations for some images (fully supervised) and image-level labels for others (weakly supervised). This reduces the annotation effort and allows the model to leverage a larger amount of weakly labeled data [34].

In comparison with traditional U-Net and other architectures, MCPAD-UNet's design allows it to take full advantage of mixed supervision. By combining both types of annotations, MCPAD-UNet can achieve competitive performance with significantly fewer and less complex annotations. This results in a cost-effective solution for defect detection in industrial steel sheets.

- MCPAD-UNet versus TLU-Net: TLU-Net (Transfer Learning U-Net) is a U-Net variant that utilizes transfer learning by initializing the encoder with pre-trained weights from a different task or dataset, such as ImageNet. This approach can be advantageous when dealing with limited annotated data, as it allows the network to leverage pre-learned features from a different domain [35].

While both MCPAD-UNet and TLU-Net leverage dilated convolutions and attention mechanisms for segmentation tasks, they differ in their architectures. MCPAD-UNet introduces HDC and DDC blocks, which enhance multi-scale feature extraction and context capture, making it more effective in handling complex images with large defect variations.

Comparing the two, MCPAD-UNet's attention mechanisms and multi-scale feature extraction capabilities make it well-suited for capturing fine-grained details and complex defect patterns in industrial steel sheets.

In summary, MCPAD-UNet demonstrates promising potential in surface defect detection compared to traditional U-Net, SegNet, and even sophisticated architectures like DeepLabv3. Its unique combination of hybrid dilation convolution, double dilated convolution, and channel and spatial attention allows it to excel in capturing details and context, making it a strong contender for automated defect segmentation in industrial settings. However, conducting a comparative study using the same evaluation metrics and datasets would provide a more concrete and quantitative assessment of the models' performance differences.

5 Conclusion

In the detection and categorization of surface defects in a variety of materials, including metals, plastics, and composites, deep learning has demonstrated encouraging results [9–11]. Because they can automatically identify characteristics from images, convolutional neural networks (CNNs) are the most often utilized kind of deep learning model for surface defect identification. Surface defect detection algorithms have performed better thanks to transfer learning [13, 14], which uses a pre-trained deep learning model as the starting point for additional training on a new dataset [30].

The limitation of widely accessible annotated training data is one of the major obstacles to surface defect identification using deep learning [31]. Researchers have utilized a variety of methods to create synthetic data or learn from unannotated data using self-supervised learning to address this. Surface defects might appear differently depending on circumstances like illumination, viewing angles, and material characteristics, which presents another difficulty. The robustness of surface defect detection models to these variations has been improved by researchers using methods like multi-view learning and domain adaptation. Overall, using deep learning to surface defect identification has the potential to increase inspection procedures' effectiveness and precision across a range of sectors, including manufacturing, construction, and transportation [27–30].

The integration of CPAM and CPAD-Net with the UNet model can bring several potential benefits for image segmentation tasks in industrial surface defect applications. Firstly, the UNet model has been proven to be effective in segmenting various types of images including medical and industrial images. By integrating CPAM and CPAD-Net, the UNet model benefits from the strengths of both models, which has led to even better performance in detecting surface defects.

The industrial surface defect detection tasks for which CPAM and CPAD-Net were specifically created have yielded outstanding results. Whereas CPAD-Net uses a cascade parallel dilated convolutional network to capture multi-scale contextual information, CPAM uses a contextual pyramid

attention module to enhance the feature representation of the model. These models can be used in MCPAD-UNet, and the resulting model can take advantage of their advantages to enhance feature representation and contextual information gathering.

Another potential benefit is the ability of the integrated model to handle complex and variable defect shapes and sizes. Industrial surface defects can come in a variety of shapes and sizes, and traditional segmentation models may struggle to accurately detect them. By integrating CPAM and CPAD-Net with UNet, the resulting model performed better at handling complex and variable defect shapes and sizes due to the improved feature representation and contextual information capture.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11760-023-02745-2.

Author contributions The author contributed to the study's conception and design. Experimental analyses were performed by AFK. The first draft of the manuscript was written by AFK and all authors commented on previous versions of the manuscript.

Funding Sakarya University of Applied Sciences BAP and TUBITAK 1505 Program supports this study with Project numbers 078-2022 and 5220125.

Data availability The data are available on request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals.

References

1. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. Comput. Vis. Pattern Recognit. **2017**, 1800–1807 (2017). https://doi.org/10.1109/CVPR.2017.195
2. Bulnes, F.G., Usamentiaga, R., Garcia, D.F., Molleda, J.: An efficient method for defect detection during the manufacturing of web materials. J. Intell. Manuf. **27**(2), 431–445 (2016). https://doi.org/10.1007/s10845-014-0876-9
3. Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder–Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Technical reports
4. Oztemel, E., Gursev, S.: Literature review of Industry 4.0 and related technologies. J. Intell. Manuf. **31**, 127–182 (2018). https://doi.org/10.1007/s10845-018-1433-8
5. Rački, D., Tomaževič, D., & Skočaj, D.: A compact convolutional neural network for textured surface anomaly detection. In: IEEE Winter Conference on Applications of Computer Vision, pp. 1331–1339 (2018). https://doi.org/10.1109/WACV.2018.00150
6. Luo, Q., Fang, X., Liu, L., Yang, C., Sun, Y.: Automated visual defect detection for flat steel surface: a survey. IEEE Trans. Instrum. Meas. **69**(3), 626–644 (2020)
7. Yu, Z., Wu, X., Gu, X.: Fully convolutional networks for surface defect inspection in industrial environment. In: International Conference on Computer Vision Systems, pp. 417–426. Springer (2017)
8. Li, Y. et al.: Research on segmentation of steel surface defect images based on improved Res-UNet network. 电子与信息学报 **44**, 1–8 (2022)
9. Jamshidi, P., Velez, M., Kastner, C., Siegmund, N., Kawthekar, P.: Transfer learning for improving model predictions in highly configurable software. In: Proceedings of the 12th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, Ser. SEAMS '17, pp. 31–41. IEEE Press, Piscataway (2017). https://doi.org/10.1109/SEAMS.2017.11
10. Su, Z., et al.: An improved U-Net method for the semantic segmentation of remote sensing images. Appl. Intell. **52**(3), 3276–3288 (2022)
11. Wu, Y., et al.: Hybrid deep learning architecture for rail surface segmentation and surface defect detection. Comput. Aided Civ. Infrastruct. Eng. **37**(2), 227–244 (2022)
12. Guo, C., Szemenyei, M., Yi, Y., Hu, Y., Wang, W., Zhou, W.: Channel attention residual U-net for retinal vessel segmentation (2020). arXiv:2004.03702
13. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA (June 2020)
14. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: a new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. **43**(2), 652–662 (2019)
15. Wang, X., Wang, S., Zhang, Z., Yin, X., Wang, T., Li, N.: CPAD-Net: contextual parallel attention and dilated network for liver tumor segmentation. Biomed. Signal Process. Control **79**(2), 104258 (2023)
16. Zhu, W., Wang, Q., Luo, L., Zhang, Y., Lu, Q., Yeh, W.-C., Liang, J.: CPAM: Cross patch attention module for complex texture tile block defect detection. Appl. Sci. **12**(23), 11959 (2022). https://doi.org/10.3390/app122311959
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, pp. 234–241. Springer International Publishing (2015)
18. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al.: Attention u-net: learning where to look for the pancreas (2018) arXiv:1804.03999
19. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation (2018). arXiv:1809.10486
20. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., Liang, J.: UNet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, pp. 3–11. Springer International Publishing (2018)
21. Sahayam, S., Nenavath, R., Jayaraman, U., Prakash, S.: Brain tumor segmentation using a hybrid multi-resolution U-Net with residual dual attention and deep supervision on MR images. Biomed. Signal Process. Control **78**, 103939 (2022)
22. Zhao, W., Chen, F., Huang, H., Li, D., Cheng, W.: A new steel defect detection algorithm based on deep learning. Comput. Intell. Neurosci. **2021**, 1–13 (2021)
23. Saiz, F.A., Alfaro, G., Barandiaran, I., Graña, M.: Generative adversarial networks to improve the robustness of visual defect

segmentation by semantic networks in manufacturing components. Appl. Sci. **11**(14), 6368 (2021)
24. Hu, J., Yan, P., Su, Y., Wu, D., Zhou, H.: A method for classification of surface defect on metal workpieces based on twin attention mechanism generative adversarial network. IEEE Sens. J. **21**(12), 13430–13441 (2021)
25. Lee, S.Y., Tama, B.A., Moon, S.J., Lee, S.: Steel surface defect diagnostics using deep convolutional neural network and class activation map. Appl. Sci. **9**(24), 5449 (2019)
26. Dong, H., Song, K., He, Y., Xu, J., Yan, Y., Meng, Q.: PGA-Net: pyramid feature fusion and global context attention network for automated surface defect detection. IEEE Trans. Industr. Inf. **16**(12), 7448–7458 (2019)
27. Severstal: Severstal: Steel Defect Detection (2019)
28. Lin, T.-Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision (ICCV), vol. 2017, pp. 2999–3007 (2017)
29. Kingma, D. P., Ba, J.L.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations, pp. 1–13 (2015)
30. Yeung, M., Sala, E., Schönlieb, C.B., Rundo, L.: Unified focal loss: generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Comput. Med. Imaging Graph. **95**, 102026 (2022)
31. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. Adv. Neural. Inf. Process. Syst. **33**, 15288–15299 (2020)
32. Yang, B., Liu, Z., Duan, G., Tan, J.: Mask2Defect: a prior knowledge-based data augmentation method for metal surface defect inspection. IEEE Trans. Industr. Inf. **18**(10), 6743–6755 (2021)
33. Urbonas, A., Raudonis, V., Maskeliūnas, R., Damaševičius, R.: Automated identification of wood veneer surface defects using a faster region-based convolutional neural network with data augmentation and transfer learning. Appl. Sci. **9**(22), 4898 (2019)
34. Božič, J., Tabernik, D., Skočaj, D.: Mixed supervision for surface-defect detection: from weakly to fully supervised learning. Comput. Ind. **129**, 103459 (2021)
35. Damacharla, P., Rao, A., Ringenberg, J., Javaid, A. Y.: TLU-Net: a deep learning approach for automatic steel surface defect detection. In: 2021 International Conference on Applied Artificial Intelligence (ICAPAI), Halden, Norway (2021)