HDR Image-Based Deep Learning Approach for Automatic Detection of Split Defects on Sheet Metal Stamping Parts

Aru Ranjan Singh[1*], Thomas Bashford-Rogers[2], Demetris Marnerides[1], Kurt Debattista[1] and Sumit Hazra[1]

[1*]Warwick Manufacturing Group, University of Warwick, Coventry, CV48AL, UK.
[2]University of the West of England, Frenchay Campus, Bristol, BR161QY, UK.

*Corresponding author(s). E-mail(s): aru-ranjan.singh@warwick.ac.uk;
Contributing authors: thomas.bashford.rogers@gmail.com; dmarnerides@pm.me; k.debattista@warwick.ac.uk; sumit.hazra@warwick.ac.uk;

Abstract

Sheet metal stamping is widely used for high-volume production. Despite the wide adoption, it can lead to defects in the manufactured components, making their quality unacceptable. Because of the variety of defects that can occur on the final product, human inspectors are frequently employed to detect them. However, they can be unreliable and costly, particularly at speeds that match the stamping rate. In this paper, we propose an automatic inspection framework for the stamping process that is based on computer vision and deep learning techniques. The low cost, remote sensing capability and simple implementation mean that it can be easily deployed in an industrial setting. A particular focus of this research is to account for the harsh lighting conditions and the highly reflective nature of products found in manufacturing environments that affect optical sensing techniques by making it difficult to capture the details of a scene. High dynamic range images can capture details of an environment in harsh lighting conditions, and in the context of this work, can capture highly reflective metals found in sheet metal stamping manufacturing. Building on this imaging technique, we propose a framework including a deep learning model to detect defects in sheet metal stamping parts. To test the framework, sheet metal 'Nakajima' samples were pressed with an industrial stamping press. Then optimally exposed, sequence of exposures, tone-mapped and high dynamic range images of the samples were used to train convolutional neural network based detectors. Analysis of the resulting models showed that high dynamic range image-based models achieved substantially higher accuracy and minimal false-positive predictions.

Keywords: HDR imaging, deep learning, sheet metal stamping, object detection, industrial inspection, defect detection

1 Introduction

Sheet metal forming processes are used primarily for high-volume products produced for a range of sectors, from white goods manufacturing to the automotive and aerospace sectors [1]. The stamping process is particularly suited for high-volume mass production (typically tens of parts per minute). As a result, despite the high investment costs of tooling (of the order of £250k per toolset), the per piece price is relatively low (compared to other processes such as casting or machining) because production costs are amortised over large volumes [2]. Furthermore, it allows the formation of complex shapes. Despite the advantages, the process can result in defects in the manufactured components making their quality unacceptable.

A variety of defects can occur due to the stamping process, with different shapes, sizes or positions, as listed in the literature [3]. The most critical defect in a stamped part is the 'split' [2]. Splits are caused when the plastic deformation that is accumulated in the material during its manufacturing exceeds its forming limits, resulting in a through-thickness fracture of the component. A split component is functionally and aesthetically unusable and is therefore scrapped. These defects are commonly seen in high-strength materials such as ultra-high-strength steels (with the ultimate tensile strength of 1000MPa) and lightweight materials such as aluminium. They have lower forming limits but are critical for automotive and aerospace components because they allow the manufacture of lightweight components that reduce transport emissions.

The most effective current technique used for detecting defects in stamped parts is by human visual inspection [4]. This method of inspection is costly, time consuming, and most importantly, prone to human errors. The impact of not detecting a defect (a false negative) is that it will proceed along the manufacturing process and will likely be assembled into a sub-assembly or the final product. Scrapping a product later in the production process incurs a greater cost and reputational damage if it is delivered to a customer. Therefore, a reliable and robust automatic defect detection method is necessary after the metal stamping process.

Several sensor-based techniques have previously been implemented for a similar task, however, sensor-based systems are unreliable and only 46% of the sensor systems placed in industries are completely functional [5]. This is because most of the sensors rely on touch, and in industrial environments touch-type sensors are prone to breaking down. However, vision-based sensors have the advantage that they do not rely on touch and utilise relatively cheap hardware.

Previously, manual techniques were implemented for feature extraction and defect detection in computer vision (CV) inspection. The manual design of the feature extractor required huge efforts for every defect type and sample shape. However, developments in deep learning (DL) especially the introduction of convolutional neural networks (CNN) have the potential to shift the paradigm from manual inspection to machine learning based object detection [6]. State-of-the-art CNN based object detectors such as YOLO [7] and Faster RCNN [8] are able to achieve >80% accuracy on a challenging large dataset (Pascal Visual Object Challenge (PASCAL VOC)) [9], while running at real-time speeds of up to 140 fps [8, 10].

In spite of the accuracy of CNN models, they have not been widely adopted by industries for sheet metal inspection. A major issue with these models is that they are unable to detect objects in extreme lighting conditions.

This work proposes a framework for automated defect detection on production lines which is designed to be robust to a wide range of illumination conditions while providing high accuracy and a low false-positive rate. We achieve this by leveraging an imaging technique known as High Dynamic Range (HDR) to capture details that conventional optical capture systems miss, then train a deep learning system to both detect and localise defects.

While HDR imaging is able to capture real-world lighting values, there are multiple ways to use this lighting information in a deep learning system. We explore these options and provide a recommendation that is able to achieve an increase of 7.2% over conventional LDR imaging. Specifically, although our approach could be trained to detect any type of defect, we evaluated the proposed framework using manufactured "Nakajima" samples containing the neck and split defects.

To summarize, the main contributions of this work are:

- A mathematical formulation was developed to identify components difficult to detect in sheet metal stamping parts.
- A framework for automated defect detection and localization for stamped metal parts which is robust to a wide range of illumination conditions.
- A deep learning based approach leads to high accuracy while minimizing the false positive rate.
- A comparison of multiple approaches using HDR imaging for defect detection.

The remainder of the article is organised as follows. Section 2 presents relevant backgrounds and literature on various associated topics. Section 3 presents a theoretical motivation for HDR imaging to demonstrate the requirement of HDR and find the component difficult to detect using traditional LDR methods. Section 4 proposed a framework. Whereas the Experiments to validate the proposed framework were presented in Section 5. Section 6 presents the experimental results and Discussion. Finally, the Conclusion is presented in section 7

2 Background and Related Work

This article includes various topics such as sheet metal stamping, HDR technology, and DL. Therefore, in this section, we are focused on providing a brief overview of various topics relevant to the study.

2.1 Sheet Metal Stamping

Stamping is a sheet metal manufacturing process, where a moving punch plastically deforms a flat sheet against a stationary die to its final shape. This deformation should be done without the sheet splitting.

Siekirk [11] identified more than 25 variables influencing the stamping process such as component geometry, material properties and process variables (friction, lubrication etc.). Despite a good process design and optimisation, the quality of stamped components (eg. the level of springback, the occurrence of splits) will vary, due to variation of materials' properties between batches [12] and the wear and tear of the tooling. Majeske and Hammett [13] studied stamping data from automobile manufacturers and found that within the same batch the part-to-part geometric variation could be as high as 30%. Similarly, Cao et al. [14] pointed out that there are variations in input variables within a batch, i.e. Variation in material strength can be as high as 20%, the strain hardening coefficient 16% and the friction coefficient 65% in sheet metals. The large number of parameters that can vary and interact with one another in a stamping process can make the process variable. Few of these parameters can be directly controlled by an operator, making the process susceptible to unexpected failures.

2.2 Deep Learning for Sheet Metal Stamping Defects Detection

The introduction of CNN and ensuing improvements have led to better accuracy and speed of detection in object detection tasks, which leads to an increase in the study of CNN based models in industrial inspections tasks [15–18]. Yang et al. [18] implemented a pre-trained CNN model to detect defects in safety vents for the power battery and achieved up to 99.56% accuracy at a 0.33% false positive (FP) rate. Further model performance was evaluated on a Raspberry Pi to indicate the framework can be implemented in a industrial setup. Cha et al. [15] proposed a CNN-based approach to detect cracks in concrete surfaces, and they compared the deep CNN (DCNN) method with two well-known traditional edge detection methods (i.e. Canny and Sobel) and found that the CNN based method was consistent in different lighting conditions compared to traditional techniques. The same team successfully designed and implemented a Faster Region-based CNN (Faster RCNN) for real-time detection of defects, including five defects: cracks in concrete, delamination of steel reinforcement, corrosion in medium and high steel and in bolts [19]. In order to implement CNN models for real-time task a *You Only Look Once* (YOLO) network was proposed by combining classification and regression task to a single end to end network [7]. Li et al. [20] improved YOLO by making it all convolutional layers, achieving a 99% detection of scratch and inscription defects in flat rolled steel surfaces at a speed of 83 FPS. Due to the end-to-end and real-time performance of the YOLO network, several researchers implemented YOLO-based models for industrial defect detection in the literature [21–23]. For example, Zhuxi et al. [21] integrated a depthwise separable convolution and parallel dual channel attention module on yolov4 to reduce the model scale and enhance the feature maps. As a result, the model results in 96.28% mAP compared to 90.47% with Yolov4 on aluminium strip surface defects. With a similar aim to reduce the model size for industrial defect detection, Zhang et al. [22] proposed a CR-YOLO network by integrating combined channel and special attention module (CBAM) to yolov4 [24]. The study also implemented the model and a segmentation model to detect defects on an edge device for real-time application. Yao et al. [23] introduced an overlapping pooling spatial attention module and a dilated convolutional module where the former module improved accuracy and reduced over-fitting, and the latter module expands the receptive field, resulting in better performance on large objects. The results showed 8.87% and 2.38% improvement in AP for the two modules respectively on area defects of light guided plates.

Following the introduction of the transformer neural network, an architecture for sequence-to-sequence tasks, many researchers tried to integrate the transformer network into CV tasks. One initial study divided the images into 16X16 patches and used them as a sequence of vectors as input to the transformer network. The method outperformed state-of-the-art networks on image classification tasks [25]. Improving on the work, Liu et al. [26] introduced a cross-connection between non-overlapping patches, achieving state-of-the-art object detection on the msCOCO dataset. Gao et al. [27] implemented a variation of the Swin transformer called “Cas-VSwin transformer” for surface defect detection. Results showed the proposed method surpassed Swin transformer and other CNN-based state-of-the-art models on SeverstalSteel [28] and NEU-DET [29] datasets. However, the transformer-based models require more data than CNN-based models. This is due to transformer-based models need to learn connection between vectors in contrast, a CNN-based model provides a well-defined connection between vectors. In the sheet metal stamping problem, gathering a large dataset is expensive, therefore CNN is a better choice for stamping defect detection.

In the context of industrial surface defect detection, Božič et al. [30] experimented on various open source datasets containing images of defects such as DAGM [31], KolektorSDD [32], and the Severstal steel defect dataset [28]. However, the images in these datasets were either artificially created or not accounting extreme lighting, moreover these datasets does not represent the the extreme lighting conditions present in stamping shops, which make object classification difficult. However, the study focuses on tackling the scarcity of annotated data by comparing weekly supervised, fully supervised and mixed models which leads to good results for samples captured in carefully controlled lighting conditions. Similarly, Shen et al. [4] proposed a CNN model based on a MobileNet architecture to detect surface defects on flat galvanised steel sheet and achieved up to 98.81% accuracy with 97 FPS. Several CNN-based studies on industrial defect detection are listed in the literature [4, 21, 23]. However, these dealt with defects in flat sheet materials, which have different types of defects and simpler geometry than stamped components most importantly not include reflective samples. To the best of our knowledge, the only study used CNN models to detect defects in stamping products is presented in literature by Block et al. [17]. The work used images containing 8845 imprint defects to train a CNN model based on RetinaNet architecture and tested on 8584 imprint defects. The study achieved a precision and recall of 90% and 92% respectively on the test dataset, which was an important development because it demonstrated the ability of the technique to detect small defects like imprint defects (with average size between 0.7 to 5mm diameter).

A significant aspect of these studies is that they were carried out on flat or non-reflective samples. As a result, the images used for training the CNN models and validating the models were not representative of the conditions encountered in an actual sheet metal stamping manufacturing environment. In more realistic conditions, lighting from overhead electric lights, sunlight from windows and factory skylights interact with the shiny metallic surface to create unpredictable specular reflections. In concave geometries, these reflections can be multiplied because of their internal reflections. Therefore the implementation of advanced imaging technique that is robust to the environment required to be explored.

2.3 HDR Technology

The dynamic range of an image is the difference between the luminance of the brightest and darkest part of an image, where the difference is measured in a logarithmic scale of base 10 ($cd/m2$). Most camera systems capture 8 bit LDR images which are limited in the range of quantised values they can represent. This is a limitation when capturing real-world lighting values which typically exhibit a significantly higher dynamic range of values. HDR capture and representation [33] mitigate this issue by increasing the bit depth available to store image values, frequently by using 16/32 bit floating-point values. Contrary to LDR, HDR images can simultaneously represent details in both dark and extremely bright regions of an image without losing information. This comes at the cost of increased memory requirements to store HDR images.

2.3.1 Tone-mapping

As HDR images are stored at a higher bit depth, they are not immediately compatible with conventional display technologies or image processing methods. Therefore, methods of mapping HDR images to conventional LDR display and processing pipelines are required, these are known as tone mapping operators (TMO). The goal of TMOs is to make minimal changes to the image while reducing the luminance from HDR to LDR, ideally preserving contrast and image details as much as possible. Typically chroma is simply quantised to 8 bits and combined with the resulting luma. A range of TMOs have been developed [33] but the best choice of TMO is frequently application specific.

2.3.2 HDR based object detection

Although a significant amount of literature exists for the technology behind HDR imaging, HDR-based object detection is rarely discussed. Rana et al. [34] studied HDR image based feature extraction using manual CV techniques and their experiments showed enhanced feature modalities compared to LDR images under different lighting conditions. To the best of our knowledge, with the exception of [35], no studies have been made of CNN using HDR images for object detection. Mukherjee et al. [35] used a 'pseudo'-HDR image dataset for the training of the CNN based models, which was created by using an expansion operator which expanded the dynamic range of LDR images. Although the model shows lower accuracy on 'pseudo'-HDR test images compared to LDR based model, a separate native HDR dataset under the extreme lighting conditions adopted from a database shows an 11% improvement in mean average precision (mAP) compared to its low dynamic counterpart.

3 Theoretical Motivation for using HDR Imaging

The automotive-grade sheet metal component has a specular appearance because their low surface roughness leads to shiny surfaces. At the same time, industries like sheet metal stamping have harsh lighting conditions depending on the power supply and weather conditions (stamping shops allow sunlight using skylights). The above reasons make capturing details in a sheet metal stamping part for defect detection difficult. Therefore, mathematical formulations are developed in the following subsections to identify the components where the object details are not captured and lead to CV-based models to struggle. We computed entropy on 120 stamping parts for the study at nine different exposure images introduced in detail later in Section 5.

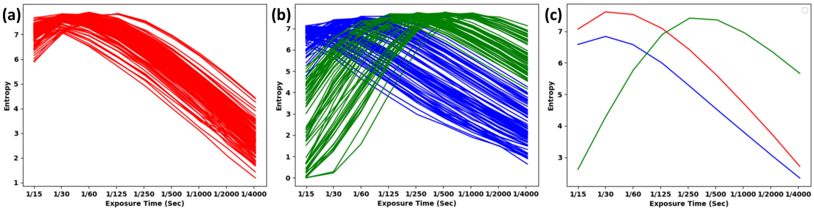

Fig. 1 Plots showing the entropy of images at a range of exposure times. These show entropy calculated on (a) the whole object (with the background removed), (b) regions of the image corresponding to defects in bright regions of the image (green) and darker regions (blue), and (c) the mean of the whole object and both defect groups.

In the field of information theory, Shannon Entropy ($H(X)$) is used to represent the amount of information in a random variable X:

$$H(X) = -\sum_{x \epsilon X} p(x) log_2(p(x)), \tag{1}$$

where $p(x)$ is the probability of a value x occurring. This can be applied to images where X is the set of intensity values, and $p(x)$ is the probability associated with the intensity value x. We computed the entropies for the whole image and areas of the image containing defects. Line plots of the entropy at different image exposure times are shown in Fig. 1. Fig. 1 (a) shows the entropy computed for the whole object (with the background removed to ensure the entropy is computed only for the part), Fig. 1 (b) shows the entropy for areas of the image containing defects in darker to well exposed regions (blue) and brighter regions (green). Fig. 1 (c) shows the mean of the whole image and the defect regions.

The study shows that captured information in an image depends on the intensity of the scene and exposure of the captured image. Similarly, depending on the intensity of light in the specific area of the scene, the captured information can vary. This analysis shows that a single exposure LDR image cannot adequately acquire the details in all image regions, even if captured at optimal exposure. Given that the location of defects is not known prior to capture and that defects may be present in both bright and dark regions of the image, an acquisition system need to capture all these details. This motivates the use of HDR imaging to capture images of defects on parts.

4 Proposed Framework

Based on the motivation (see section 1 and 2) a reliable early detection of defective components is essential for most manufacturing processes. The inspection of defective parts has to balance defect detection when present, yet minimise the number of defect-free parts flagged as containing defects, known as false positives. Minimising false positives is also essential from a financial and time perspective; if too many parts are incorrectly classified then this too many

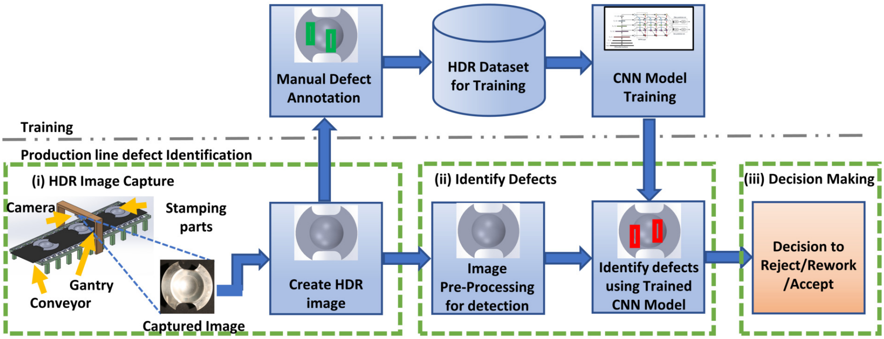

Fig. 2 A high-level overview of HDR based CNN model implementation to detect defects in sheet metal stamping assembly line. The upper part of the image (above the dash-dot line) shows the training of the HDR based CNN model and the bottom part shows the implementation of the model in the production line. In the figure, (i) HDR image capture, (ii) Identify Defects, and (iii) Decision making: Finally, using the information of defects in a sample, the decision can be made to accept/reject or rework a sample.

lead to financial losses and interruptions to the assembly line. Therefore, automated approaches to defect detection must consider both these criteria. In this section we propose a framework whereby components on a stamping line are first imaged with HDR cameras, then these images are passed through a trained deep neural network which both identifies and localises defects if present. The use of HDR imaging enables details that are lost in LDR images to be passed to the neural network, which enables the network to reliably detect the defects.

Our framework (see Fig. 2) consists of two broad steps. In the first step, a training step, HDR images of defective and good parts are captured. These images are used to train a CNN model. The second step is an online process that is used directly on the assembly line. It involves a number of stages:

- HDR image capture: The stamping parts move to assembly points through conveyors. In mass production, to stop the defective parts from reaching an assembly point, the defects need to be identified directly on the conveyors. The framework proposes capturing HDR images of the parts directly from a gantry over the conveyors.
- Identifying defects: In this step, the HDR images are preprocessed as required by the model. Then the model will detect the defects.
- Decision on the defective part: Finally, in this stage depending on the size, position and type of defects a decision can be made to accept/ reject or to send for repair.

An overview of our framework is shown in Fig. 2, and each of these steps is discussed in more detail below.

4.1 HDR image capture

The first stage of both the training and online process is to capture HDR images. There are two common methods to achieve this: either using native

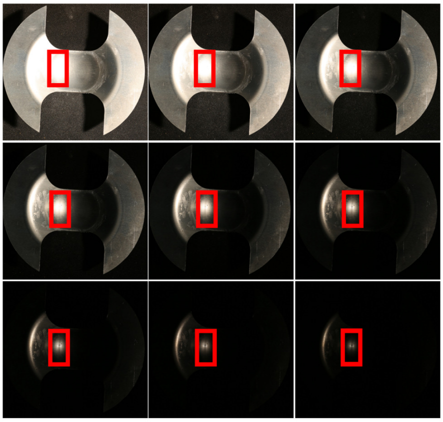

Fig. 3 An example of nine exposure images extracted from a single HDR image. The exposure time was reduced in reading order. Row 1 images are the higher exposure images where details at less bright regions are captured. Row 3 are lower exposure images where the defect details in the brightest spots are captured with a cost of losing details in less bright regions.

HDR cameras or using bracketing techniques. Native HDR cameras are capable of capturing scene details with high dynamic range in a single shot. However, these native HDR cameras are not widely available compared to general purpose cameras. The bracketing technique uses general purpose cameras to capture a series of images of the scene at different exposures [33]. These different exposures are typically achieved by changing the exposure time while keeping the ISO and aperture constant. The various exposure times allow the camera sensor to collect a different number of photons each time per pixel. This results in a range of images being captured including under and over exposed images. This set of images are then merged to create an HDR image of the scene. Various merging algorithms have been proposed to merge the images [33], among all the most common is the Debevec merging algorithm [36]. Regardless of the method of capture, the output from this stage results in an HDR image, I_{HDR}, of the part.

This is able to represent the details of a part regardless of the scene illumination and is capable of storing details of defects even in very bright illumination conditions. Fig. 3 illustrates this with a capture of a real part with a defect which happens to fall in a bright reflection. This image shows nine exposures extracted from a single HDR image, where details in the darker regions are visible on the top row of higher exposures, but the defect is clearly visible on the bottom row of lower exposures. The HDR image encodes all this information.

4.2 CNN Model

A neural network object detector is also required for both training and online detection. This requires a choice of the network model. The state-of-the-art object detection models can be classified into two categories: a) two stage or region proposal networks (RPN) based algorithms, and b) single stage detectors. The RPN-based models first generate region proposals followed by object detection. Whereas, the single stage detectors are based on global classification and regression which can predict class probability scores and bounding box locations directly from the input images using a single feed forward CNN model. Since the model is a single feed forward network the model design is relatively straightforward and can train and optimise end-to-end. State-of-the-art models in this category include YOLO, which is capable of detecting in real-time [7]. Although the framework can be implemented on both types of models in this study a single stage detector is implemented.

After the creation of datasets, the images were processed through a CNN network for training and testing. Normally a CNN network takes an image $I \in \mathbb{R}^{w \times h \times c}$ as input and predicts the object classes and locations with a confidence value. Commonly, CNN models are designed and trained on LDR images with a single image as an input. For this framework, we can use any backbone of choice but add custom inputs to the second layer while leaving the rest of the network fixed. We represent the fixed part of the backbone as $y_{cnn,2:}$ which is connected to a representation dependent input through a function $f(\cdot)$ as defined below. Therefore, the prediction from the image can be written as:

$$\mathbf{V} = y_{cnn,2:}(f(I)) \tag{2}$$

where the function $f : \mathbb{R}^{w \times h \times c} \mapsto \mathbb{R}^{F}$ maps the input image or set of images I through a differentiable transform to the tensor encoding of the second layer of the network which is of size F. This is required to work with different representations of HDR images as discussed in Section 5.3.

4.3 Adapting HDR images as CNN inputs

While we propose the use of HDR images as inputs to a CNN detector, this is not straightforward, as CNNs are usually designed for use with LDR images. There can be difficulties in directly using HDR content, for example if leveraging pre-trained networks that require inputs in the range [0..1] (HDR images are unbounded $I_{\text{HDR}} \in \mathbb{R}^{+}$), that were trained using LDR image statistics. Even if training from scratch, HDR image statistics can be hard to normalise, as is usually done for CNN inputs. Therefore, the function $f(\cdot)$, defined above, must also implement a mapping which encodes the HDR image in the range [0..1] and appropriately adapts the image statistics. We, therefore, propose, as a final part of our framework, three methods for adapting HDR images to be used as inputs to CNNs. We analyse their performance in the next section.

The first approach is to extract one or more exposures from the HDR image similar to Fig. 3. If one exposure is used, then the optimal exposure can be chosen which is equivalent to capture one single exposure image at optimal exposure.

$$I_0 = I^e_{\mathrm{HDR}}, \text{where } e = \underset{\hat{e}}{\operatorname{argmax}} \sum_i^{w*h} 1(t_l < I_{\mathrm{HDR}}(i)^{\hat{e}} < t_u) \qquad (3)$$

where the indicator function 1 returns 1 if a pixel value is between a lower t_l and upper threshold t_u and 0 otherwise.

Multiple exposures are simply extracted from computing the optimal exposure, then taking fixed offsets. These can then be concatenated when input to the network.

A second approach is to map the HDR values to LDR values through a process known as tone mapping. A Tone Mapping Operator (TMO) takes I_{HDR} as input and returns a tone-mapped image which is in the range [0..1] while preserving as many details as possible. Usually, the TMOs only reduce the luminance range, while colours are unprocessed.

$$\mathrm{TMO}(I) = \begin{cases} L_d = f_l(L_w) : R_i^{w \times h} \rightarrow [0..1] \\ \begin{bmatrix} R_d \\ G_d \\ B_d \end{bmatrix} = L_d \left(\frac{1}{L_w} \begin{bmatrix} R_w \\ G_w \\ B_w \end{bmatrix} \right)^s \end{cases} \qquad (4)$$

where $s \in [0, 1]$ is a saturation factor that decreases saturation. After the application of tone mapping operator TMO(·), gamma correction is usually applied to each colour channel, followed by clamping to [0..1]. Many TMOs exist, each designed for different image types. We discuss this more in Section 5.3.

The final method, directly normalises the HDR image by dividing by the maximum value: $\frac{I_{\mathrm{HDR}}}{\max(I_{\mathrm{HDR}})}$. This does preserve all details, but if the HDR image contains outliers then the majority of the image may contain extremely small values which can cause difficulties when training [33].

5 Experiments

In this section, we describe a comparison study conducted to evaluate the performance of the proposed framework. This includes (i) sample preparation: describes the manufacturing of the defective and good samples, (ii) datasets creation: describes the image capture and object annotation for creation of different image datasets, and (iii) implementation details.

5.1 Sample Preparation

This study used the Nakajima samples that are used in standard forming limit curve (FLC) tests (see Fig. 3) as outlined in [37]. In order to manufacture the samples first, raw sheet materials were cut to the correct shape and size using a Datron CNC milling machine. As this work is interested in the defects, this method did not follow the exact sets required for an FLC evaluation. Instead, hemispherical samples of different sizes were randomly generated. Then using an Interlaken 225 press, samples were pressed to their final shape. In order to generate neck or split defects, the samples were deformed until load drops associated with the formation of necks or splits were detected. The main aim of this work was to generate neck or splits, but the samples also included other defects like wrinkles, edge cracks and scratches that were formed as an inevitable part of the tests. In addition, scratches were also formed during the handling of samples. In contrast, non-defective parts were produced by pressing samples to a punch height prior to necking or splitting. This punch height was determined through trial and error. The study produced 120 samples containing 104 splits and 40 necks divided into five-fold cross-validation sets. An additional 30 samples without necks or splits were used as safe samples only during the model validation. These samples also contain other defect types.

5.2 Datasets Creation

The proposed framework was evaluated by comparing the CNN models trained on five different datasets to test the HDR mappings as discussed in Section 4.1. Specifically, we generated and tested a single optimal exposure image, a set of images at lower exposures designed to capture details in brighter regions (offsets of -3 and -5 from the optimal value), two commonly used tone mapping operators: the Reinhard local operator [38] (ReinhardTMO) and the Ward histogram adjustment operator [39] (WardHistAdjTMO), and the use of normalised HDR images.

Images for this experiment were captured with the following setup:

5.2.1 Equipment

In this study, the images were captured in a dedicated image capture room, with no external light intervention. To illuminate the scene a 2,000W lamp was used and the camera adopted to capture a sequence of nine exposures was Canon EOS 5D Mark III in jpeg format.

5.2.2 Procedure

The HDR image's purpose is to capture the scene's full dynamic range. Therefore, in the bracketing technique, the highest and lowest exposure should capture the details in the darkest and brightest spots in the scene. Furthermore, the noise in the final image after merging reduces with the number of in-between exposures since the merging algorithm compensates for the missing information. Therefore it is a common practice to capture images every twice exposure covering the dynamic range to create an HDR image. From Fig. 1, it can be seen that the peak of the information for splits varies between exposure time 1/15 to 1/4000 at constant ISO and aperture. Therefore in our study, we captured a sequence of images at every exposure time twice from 1/4000 seconds to 1/15 seconds, when keeping the ISO and aperture constant.

5.2.3 Image processing

The captured nine different exposure images were merged together using Debevec merging algorithm from OpenCV 4.4.0 to create HDR images. Then the tone-mapped LDR images were created using ReinhardTMO and WardHistAdjTMO from MATLAB HDR toolbox [33], we will discus this more in section 5.3. Finally, the resulting images cropped to 3100 × 3100 then resized to 1024 × 1024 pixels.

5.2.4 Image Annotation

The neck and split defects were annotated as a single class since they have similar characteristics. The annotation was created manually using a graphical image annotation tool LabelImg [40]. "Each sample was carefully observed for any physical defects as a reference then the optimal-exposure image was opened in the LabelImg tool and annotation boxes were created surrounding the area of the visible defects." Further, the other exposures and tone-mapped images of the same sample were opened one by one and the annotation boxes were refined.

5.3 Implementation Details

This study used yolov5: v3.1 as backbone $y_{\text{CNN},2:}$, see section 4.2. The latest version of the YOLO series is a state of the art CNN architecture for image classification and object detection [10]. We describe the mapping functions $f(\cdot)$ related to the representations of HDR images introduced in section 4.2 below.

To define the mapping functions from the image to the backbone, we first define some common transformations. In YOLOv5 F (see section 4.2) is of size $(64 \times w/2 \times h/2)$ and all our functions eventually map to this dimension. Normally in backbones, a higher kernel size with a stride greater than one or a pooling layer is used at the first layer to reduce the resolution by increasing the depth. The same task is accomplished by a SpaceToDepth stem layer (i.e. focus layer in yolov5) y_{focus} introduced by [41] at a low computational cost. The Focus layer rearranges the block of spatial data to depth, which reduces the resolution. Therefore using a smaller kernel, can effectively convolve on a higher number of pixels. The Focus layer converts images dimension (c, w, h) to $(4 \times c, w/2, h/2)$, where and c is the number of channels of the input.

Therefore the function for optimal-exposure images is

$$f_{BE}(I_0) = conv3\left(y_{focus}\left(\frac{I_0}{255}\right)\right), \tag{5}$$

where $conv3 : R^{12\times w/2\times h/2} \mapsto \mathbb{R}^{32\times w/2\times h/2}$ is a convolution kernel which maps the result of y_{focus} to a tensor with higher depth, thereby learning low level features of the image before input to the network.

The function for 3-images is:

$$f_{3E}(\{I_0, I_{-3}, I_{-5}\}) = conv3_E\left(y_{focus}\left(\frac{cat(\{I_0, I_{-3}, I_{-5}\}}{255}\right)\right), \tag{6}$$

where $I_{3E} = cat(I_0, I_{-3}, I_{-5})$ concatenates three single exposure images channel wise and $conv3_E : R^{36\times w/2\times h/2} \mapsto \mathbb{R}^{64\times w/2\times h/2}$ maps from the concatenated images to the second layer of the backbone. This number of features is equal to the size of the second layer of YOLO, i.e. equal to F and is designed to fuse information from the three single exposure images.

The functions for tone-mapping images use the Reinhard and Ward operators. The ReinhardTMO is based on photographic principles which simulate burn and dodge effects. The operator is be defined as

$$L_d(\mathbf{x}) = \frac{L_m(\mathbf{x})\left(1 + L_{white}^{-2} L_m(\mathbf{x})\right)}{1 + L_{\sigma_{max}}(\mathbf{x})} \tag{7}$$

where $L_m(\mathbf{x})$ is original luminance at pixel index $\mathbf{x}$ scaled by $aL_{w,H}^{-1}$ and a is the chosen exposure can be automatically estimated [42]. $L_{w,H}$ is the logarithmic average of the luminance. L_{white} is the smallest luminance value mapped to white, i.e. equal to $L_{w,max}$ by defult. If $L_{white} < L_{m,max}$, values greater than L_{white} are clamped. Finally $L_{\sigma_{max}}(\mathbf{x})$ is the average luminance computed over the largest neighbourhood (σ_{max}) around the image pixel. In our context, we define this operator as $TMO_R(I_{\text{HDR}})$, and is used as follows:

$$f_R(I_{\text{HDR}}) = conv3\left(y_{focus}\left(\frac{TMO_R(I_{\text{HDR}})}{255}\right)\right) \tag{8}$$

The Ward histogram operator operates on the histogram of HDR values. This first truncates the histogram $H(\mathbf{x})$ without changing the total number of samples T, then based on the cumulative frequency a histogram equalisation formula is implemented to compute the LDR luminance:

$$L_d(x) = \exp\left(\ln(L_{d,min}) + P(\ln L_w(x)) ln(L_{d,max}/L_{d,min})\right) \tag{9}$$

where $L_{D,min}$ and $L_{D,max}$ respectively minimum and maximum luminance of tone-mapped image. P is the cumulative histogram. During the histogram adjustment higher contrast is reduced by using a linear ceiling on the contrast, produced by TMO (i.e. the contrast in any given region should not exceed contrast produced by the TMO).

$$\frac{dL_d(\mathbf{x})}{dL_w(\mathbf{x})} \leq \frac{L_d(\mathbf{x})}{L_w(\mathbf{x})}. \tag{10}$$

The iterative process to truncate the histogram continues until it satisfies this condition, derived by taking derivative of eq. 5.3 and followed by applying the linear ceiling (eq. 10).

$$H(x) \leq \frac{T\Delta x}{\ln(L_{d,max}/L_{d,min})}. \tag{11}$$

Similar to Reinhard, we apply this to the network using $TMO_W(I_{\text{HDR}})$ to represent the WardHistTMO.

$$f_W(I_{\text{HDR}}) = conv3\left(y_{focus}\left(\frac{TMO_W(I_{\text{HDR}})}{255}\right)\right) \tag{12}$$

Finally for HDR images, the following mapping function is used:

$$f_{\text{HDR}}(I_{\text{HDR}}) = conv3\left(y_{focus}\left(\frac{I_{\text{HDR}}}{max(I_{\text{HDR}})}\right)\right) \tag{13}$$

which normalises the input HDR image by its maximum value, i.e. ensures that regardless of the dynamic range of the image, the network receives input between 0 and 1 as expected.

Originally the Yolov5 model is designed and trained for LDR images. During training mosaic data augmentation as well as horizontal, vertical flipping, scaling, shifting, and colour space augmentation were applied. All the augmentation were replicated during the training for all the datasets. The images were normalised before proceed to the Yolov5 model which makes the cost of computation the same for all image types. The models were trained from the scratch and run for 2000 epochs with a batch size of 16. The hyperparameters were kept as proposed in the original resource [10] without any fine tuning. The device used for testing and training the models used an Intel I9-9980XE CPU and an NVIDIA Quadro RTX 5000 GPU, running Ubuntu 18.04+ as an operating system using CUDA 10.1, python3.7 and torch 1.4.0 libraries.

6 Results and Discussion

6.1 Results

The performance comparison of models using different datasets is shown in Table. 1. The table lists mAP50 (mAP at the intersection over union (IOU) threshold 0.5) and mAP (average of mAP at IOU thresholds 0.5-0.95 at an interval of 0.05) for each fold and average of the folds.

The average of mAP50 and mAP for folds shows improvement of 7.4% and 4.8% respectively for the HDR image based model compared to the single optimal-exposure image based model. Although the mAP50 reaches 98.9%, the mAP is below 67.7% for all the models this is because the mAP90 and mAP95 (mAP at IOU threshold 0.90 and 0.95 respectively) were low i.e. up to 0.106 and 0.007 respectively for the models. In object detection, correct and tight annotation is always an issue whereas, specific to the problem splits start and vanish gradually without any clear edge which causes the length of the splits to vary depending on the annotator.

For the further evaluation of the performance IOU threshold of 0.5 was selected since in the sheet metal stamping defect detection task cost involved in detecting a defect is significantly higher than the cost involved in the precise detection. Using 0.5 IOU threshold at different confidence threshold the true positive (TP: a defect predicted as defect with having IOU greater than

Table 1 AP at IOU 0.5 and mAP at IOU 0.5-0.95 in an interval of 0.05 listed for each fold and average of the folds

	mAP50						mAP					
Dataset	Fold1	Fold2	Fold3	Fold4	Fold5	Avg	Fold1	Fold2	Fold3	Fold4	Fold5	Avg
Optimal-Exposure	0.918	0.857	0.902	0.954	0.944	0.915	0.648	0.567	0.629	0.630	0.654	0.626
3-images	0.953	0.949	0.991	0.996	0.995	0.977	0.693	0.590	0.650	0.666	0.660	0.652
ReinhardTMO	0.957	0.948	0.990	0.981	0.975	0.970	0.711	0.590	0.648	0.666	0.660	0.655
WardHistAdjTMO	0.947	0.986	0.967	0.966	0.964	0.966	0.650	0.588	0.598	0.645	0.637	0.624
HDR	0.967	0.994	0.994	0.996	0.996	**0.989**	0.719	0.583	0.672	0.686	0.709	**0.674**

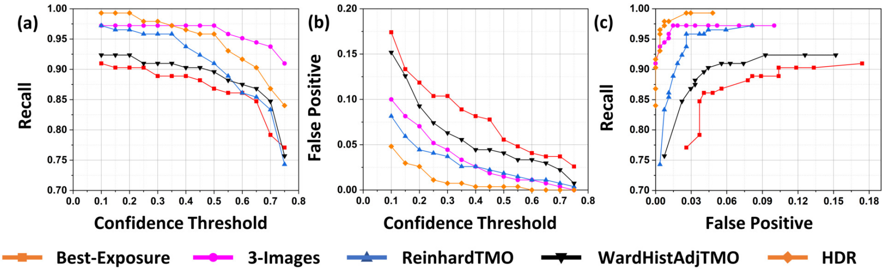

Fig. 4 Shows (a), and (b) Recall and False positive plots respectively with respect to confidence threshold, and (c) receiver operating characteristic (ROC) curve at IOU threshold of 0.5.

Table 2 Methods ranked based on mean accurate prediction with combined Cochran's Q-value and the groups represent where there are no significant difference between methods

Models with Mean Ranking					Q-Value	Sign
HDR (0.993)	3_images (0.972)	ReinhardTMO (0.965)	WardHistAdjTMO (0.924)	LDR (0.903)	27.8049	P<0.0001

a threshold), false positive (FP: good regions predicted as defects), and false negative (FN: the defects are not predicted) were evaluated for each fold and added to calculate the total for the models. Further, the recall, FP at different confidence thresholds and ROC curve were estimated, where a significant improvement for our proposed models was observed as shown in Fig. 4.

To further verify the significance of the models performance, Cochran's Q test was conducted for all the defects and predictions from model with a confidence higher that 0.2. To carry out the study for each defect, overlapping of predictions were calculated and and IOU greater than 0.5 is considered as "1" (correct prediction) otherwise "0" (not correctly predicted). For the combined test the Q-value (27.8) is larger than Q-Critical (9.5) at 0.95 confidence level and 4 degrees of freedom. Further post-hoc pair wise Q-tests were carried out. The statistical test results are shown in Table. 2. The methods are grouped together has no significance difference are found, however there are significant difference between separate groups. For example except WardHistAdjTMO all other HDR based methods has significant difference from traditional LDR based model.

Additionally Chi-square hypothesis test were conducted on TP and FP results for various confidence thresholds comparing HDR and HDR representation based models with LDR based model. Chi-square hypothesis tests at the 0.95 confidence level were conducted to examine the significance of the difference. From the chi-square test, it was found that, the improvement for WardHistAdjTMO based model is not significant, whereas ReinhardTMO

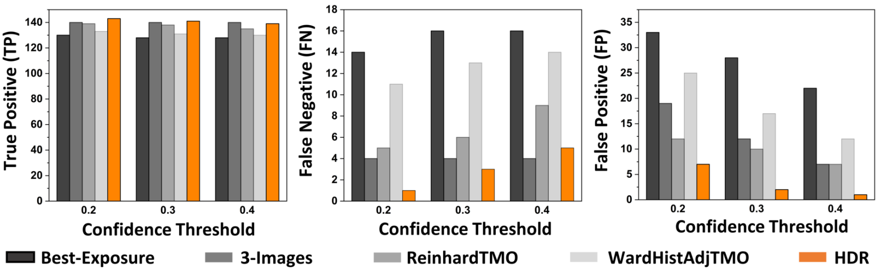

Fig. 5 Shows column plots of (a) true positives, (b) false negatives, and (c) false positives at confidence thresholds 0.2, 0.3 and 0.4 with IOU 0.5.

Table 3 Time taken for data loading and inference of methods. The time shows in seconds

Task	Best-exposure	3-images	ReinhardTMO	WardHistAdjTMO	HDR
Dataload	0.000016	0.000016	0.000016	0.000016	0.000016
Inference	0.006976	0.008118	0.006902	0.006846	0.006771

showed significant improvement of recall at 0.1, 0.15, 0.2, 0.3 and 0.35 confidence thresholds and significant reduction of FP at 0.1, 0.2, 0.3, 0.35, 0.4, and 0.45 confidence thresholds. The 3-images based model showed significant improvement in recall at all confidence thresholds, but for the reduction in false positives, the model was significant at 0.35, 0.4, 0.45, 0.5, 0.65, 0.7, 0.75 confidence thresholds. Finally, our proposed HDR model shows a significant improvement for recall as well as a significant reduction of FPs at all confidence thresholds.

Although a higher recall for the 3-image dataset was observed above the 0.4 confidence threshold, the highest recall achieved by the model was 0.972. The same recall was achieved by HDR models with lower false positives which can be observed from the ROC curve (Fig. 4 (a)). Further for the clarity of the results column plots showing TP, FN, and FP at 0.2, 0.3 and 0.4 confidence thresholds are shown in Fig. 5. From the 144 splits in the dataset, TP shows correctly detected splits and FN shows splits which are not detected. Whereas, FP shows the predictions which are incorrect. From Fig. 4 and Fig. 5 it is clear that the HDR based model not only improves the accuracy of prediction but also reduces the number of incorrect predictions, which is highly beneficial as outlined in section 4.

The data loading and inference time were computed for all the methods as presented in Table. 3. In the study, the HDR images are processed to adapt as input to the CNN model. Therefore the time required for data loading and model inference is compared. Except for the 3-images model, all other model takes similar time for inference. In contrast, all image types took similar data loading times since, irrespective of image type, the images are normalised to [0, 1] before passing through the CNN layers.

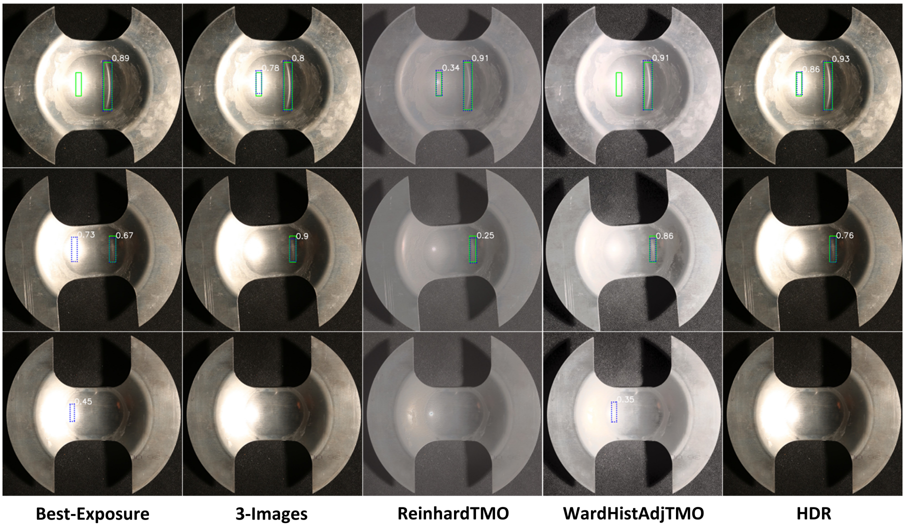

Fig. 6 Qualitative results: showing prediction (dotted blue box) and ground truth (solid green box) on samples of all datasets at confidence threshold 0.2 and IOU threshold 0.5. (Row 1 and Row 2) for test set with FN in optimal-exposure and WardHistAdjTMO image and FP in the optimal-exposure image, and (Row 3) shows FP on good samples

6.2 Visualisation of Detection

Fig. 6 shows qualitative results from the models on the test images at a confidence threshold of 0.2, where row 1 shows examples of FN from the optimal-exposure and WardHistAdjTMO image based models, and row 2 shows FP predictions for the optimal-exposure image based model. Most of the FN and FP in optimal-exposure images were observed in the higher brightness regions of the image as expected. Our proposed use of HDR imaging is able to overcome this limitation.

The FN and FP prediction from HDR dataset at 0.2 confidence threshold for 0.5 IOU threshold are shown in Fig. 7. From Fig. 5 (c) the HDR based model shows 6 FP compared to 32 FP for traditional LDR based methods at 0.2 confidence whereas increasing the confidence threshold to 0.3 reduced the FP to 2. The only FN found on the dataset shown in Fig. 7 (a) was analysed and found that for the sample the thickness reduction was mild and the reduction can only be observed on the rear side of the sample which is rare in the dataset. This can be solved with a higher number of similar samples followed by weighting the dataset to balance the defect type. Fig. 7 (b, c, and d) shows false predictions in the defective test set and Fig. 7 (e, f, and g) shows the false predictions on good samples. From the figure, it can be seen that all of the false predictions came from the scratch marks on the samples or from multiple predictions of the same defects. Increasing the number of scratch data in training can further reduce the wrong prediction.

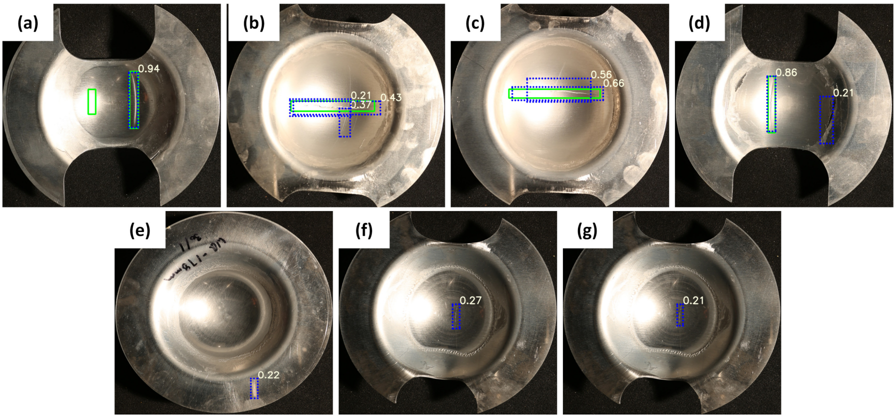

Fig. 7 Shows the missed and false detection from HDR test set, where ground truth is depicted in solid green box and predictions are in dotted blue box. In figure (a) depict FN, (b), (c), (d) depicts FP in defective samples, (e), (f), and(g) depicts FP in good samples.

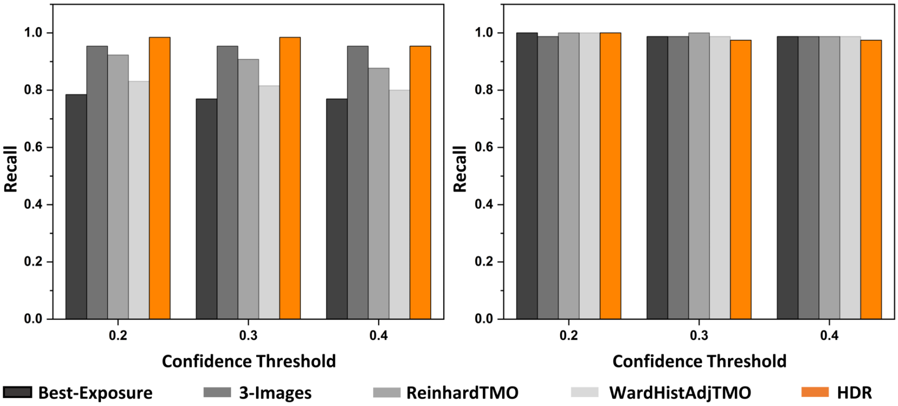

Fig. 8 shows column plots for recall of (a) defect group green, and (b) defect group blue.

6.3 Effect of Intensity of Light

As discussed in section 3, the captured information can vary depending on the crack position and light reflection. Therefore, in this section, we compare the results from HDR and the other HDR representation-based models with traditional LDR-based model results for both the group of defects as clustered in section 3. Fig. 8 shows column plots for recall for both groups of defects at 0.2, 0.3 and 0.4 confidence thresholds. As observed in Fig. 1 (b), all model recalls are comparable for this group (blue) of defects. However, the group of defects in the reflection region showed 20% improvement in recall. Whereas other HDR representation models showed 17%, 14% and 4.6% improvement for the sequence of images, ReinhardTOM and WardHistAdjTMO, respectively (see Fig. 1 (a)). From the results, it is clear that most of the improvement of our proposed method comes from defects that fall under the reflection area.

Table 4 Distribution of defects in Dataset

Defects in Sample	Number of Samples	Number of Splits	Number of Necks
Splits Only	80	80	0
Necks Only	16	0	16
Necks and Splits	24	24	24
Total	120	104	40

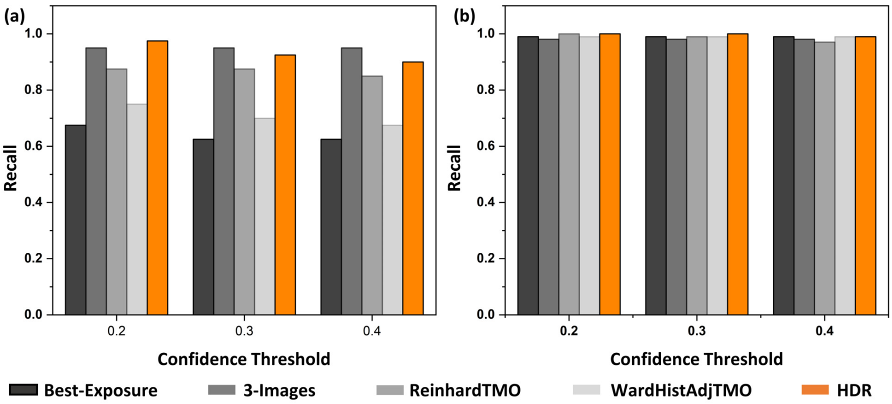

Fig. 9 shows column plots for recall of (a) necks, and (b) splits.

6.4 Effect of Geometry

Similar to the reflection of light, defects' geometry can affect the result matrix. Therefore, in the study, defects are divided into two subsets: necks and splits, where splits are cracks that split the sample, and necks are local thinning. In most studies, the experimenter divides the necks and splits them visually. However, in this study, the defects are divided into necks and splits based on there variance and location of crack. The defect is labelled as splits when the variance is above a threshold. Otherwise, they are labelled as necks. However, the variance can be less when the defect is present in a bright scene area. Therefore, the optimal exposure image pixels were first clustered into bright and normal pixels using K-means clustering based on their intensity values. Then, depending on the location of the defects in the image, the defects are assigned as bright or normal defects. Further variance thresholds were determined by correlating variance with physical samples. The final table of the neck and splits defects are shown in table 4, and a comparative result for recall of both groups are presented in Fig. 9. From Fig. 9, recall for the splits are comparable, whereas for necks, the HDR-based model improves recall by 20% compared to the traditional LDR-based model. Similarly, other HDR variations achieved considerable improvement in recall, as observed from Fig. 9.

6.5 Discussion and Limitations

The study's main aim was to propose an HDR-based framework that performs better in harsh lighting conditions. To achieve better model performance requires the ability to extract relevant information from high contrast images from a typical manufacturing scene. In section 3 we show that maximum information of the presence of a defect in a whole image and a partial part of image can appear at separate exposure. The appropriate exposure level depends on the location of the defect on the inspected part with respect to the location and intensity of a reflection from a light source. This finding is the basis of our hypothesis/motivation that HDR imaging is necessary for inspection of sheet metal components in an industrial environment (see Fig. 1). The results (see Table 1) and subsequent statistical study show that using HDR images with CNN models can improve the detection recall and reduce the number of false predictions compared to LDR-based models. In particular, our HDR-based framework performs better at detecting defects where the details are not captured by a traditional LDR image and difficult to detect objects such as necks, as shown in Fig. 8 and Fig. 9 respectively. Moreover, the study also implemented various HDR representations such as the sequence of images, and tone-mapped images can improve performance over LDR images. However, from Table 2, HDR images can beat all other methods significantly except methods using the sequence of images. However, the sequence of images (3-images) takes more processing time compared to other methods.

The proposed framework is not limited to split or neck classes of defects. Depending on the industrial requirement, this framework can detect other defects, such as wrinkles, edge cracks and scratches in stamping parts where extreme lighting conditions are a concern. A CNN model using the proposed framework can be trained to detect other classes of defects using a dataset containing sufficient labelled examples of the defects class of interest.

In this study, the HDR-based YOLOv5 model was used. In the framework, the HDR data are processed to adapt as input to the CNN model without changing the model architecture. Therefore the framework can be easily integrated with other CNN-based models with minor changes. Further, the data loading time and inference time are computed as presented in Table 3. The data loading and inference time for HDR and LDR images are similar because, irrespective of the image type, the data are normalised between [0, 1] before passing through the CNN model. Although capturing multiple images can increase the time of inspection for HDR images. However, the specialised HDR camera can complete the task in a single shot. Several industries using human vision and CV-based inspection could benefit from our research, particularly inspections struggling with extreme lighting conditions.

Although the results show in Section 6 are promising and improvements in defects detection are indeed possible using HDR image based CNN models, there are several limitations to the current framework. A major limitation of using the technique on an assembly line where the time available for detection is constrained. Taking multiple exposure images using the same camera takes few seconds. This however can be solved by using native HDR cameras. Secondly, image annotation for HDR imaging is not easy to do as there is no software available for HDR image annotation. Another limitation is there are no large HDR datasets for object detection available compared to its LDR counterpart such as PASCAL VOC, and msCOCO. This is an area we intend to tackle as future work. Finally, there are few limitations native to this work. Fine-tuning, and anchor box optimisation were not carried out in the study, which improves the accuracy of object detection models significantly. Another limitation was that a total of 150 sample images were used in the study including 30 good samples which include one object class and single lighting condition. Further testing using higher number of samples with several sheet metal stamping parts, lighting conditions, and defect classes can determine the efficacy of the framework in higher accuracy.

7 Conclusion

In this article, we proposed a framework using HDR imaging and deep learning to automate defect detection under a wide range of illumination conditions. This is suitable for production lines which require high speed and accuracy of detection alongside minimal false positive rates. To validate our approach, we studied split and neck defect detection on “Nakajima” or waisted-geometry test samples. However, this method can also be used to detect other classes of defects in sheet metal stamping parts and easily be adopted into existing or future deep learning frameworks. We investigated four representations of HDR data, including native HDR values, tone mapped images, exposure sequences and optimally exposed images. Native HDR imaging showed a significant improvement of 7.4%, and tone mapped images also exhibited a 5.1 - 5.5% improvement over using LDR images, alongside a substantial drop in the false positive rate. We believe that our approach is a step forward in reliable automated defect detection and we hope that this type of approach will be adopted in production contexts.

Declarations

Authors contribution

Material preparation was performed by AS and SH. Data collection and analysis were performed by AS, TB, DM, KD, SH. The frst draft of the manuscript was written by AS. TB, KD, SH commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

This research was financially supported by the Warwick Manufacturing Group of University of Warwick. Award Number: WIIT, received by Aru Ranjan Singh.

Availability of data and materials

The datasets used or analysed during the current study are available from the corresponding author on reasonable request.

Code availability

The code used during the current study are available from the corresponding author on reasonable request.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing Interests

The authors declare no competing interests.

References

[1] Small, N.: A statistical method for determining and representing formability innovation report. Thesis (2015)

[2] Wang, H., Liu, L., Wang, H., Zhou, J.: Control of defects in deep drawing of tailor-welded blanks for complex shape automotive panel (2021)

[3] Ghosh, S.: Principally on sheet metal forming defects as described in the eleventh biennial congress of the international deep drawing research group (iddrg). International Journal of Mechanical Sciences **23**(4), 195–211 (1981)

[4] Shen, Y., Sun, H., Xu, X., Zhou, J.: Detection and positioning of surface defects on galvanized sheet based on improved mobilenet v2. In: 2019 Chinese Control Conference (CCC), pp. 8450–8454 (2019)

[5] Garcia, C.: Artificial intelligence applied to automatic supervision, diagnosis and control in sheet metal stamping processes. Journal of Materials Processing Technology **164**, 1351–1357 (2005)

[6] Andreopoulos, A., Tsotsos, J.K.: 50 years of object recognition: Directions forward. Computer vision and image understanding **117**(8), 827–891 (2013)

[7] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)

[8] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)

[9] Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1), 98–136 (2015)

[10] Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., Laughing, tkianai, Hogan, A., lorenzomammana, yxNONG, AlexWang1900, Diaconu, L., Marc, wanghaoyang0106, ml5ah, Doug, Ingham, F., Frederik, Guilhen, Hatovix, Poznanski, J., Fang, J., Yu, L., changyu98, Wang, M., Gupta, N., Akhtar, O., PetrDvoracek, Rai, P.: ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements (2020). https://doi.org/10.5281/zenodo.4154370

[11] Siekirk, J.F.: Process variable effects on sheet metal quality. Journal of Applied Metalworking **4**(3), 262–269 (1986)

[12] Zhang, W.: Design for uncertainties of sheet metal forming process. PhD thesis, The Ohio State University (2007)

[13] Majeske, K.D., Hammett, P.C.: Identifying sources of variation in sheet metal stamping. International Journal of Flexible Manufacturing Systems **15**(1), 5–18 (2003)

[14] Cao, J., Kinsey, B.L., Yao, H., Viswanathan, V., Song, N.: Next generation stamping dies—controllability and flexibility. Robotics and Computer-Integrated Manufacturing **17**(1-2), 49–56 (2001)

[15] Cha, Y., Choi, W., Büyüköztürk, O.: Deep learning-based crack damage detection using convolutional neural networks. Computer-Aided Civil and Infrastructure Engineering **32**(5), 361–378 (2017)

[16] Li, W.-b., Lu, C.-h., Zhang, J.-c.: A local annular contrast based real-time inspection algorithm for steel bar surface defects. Applied Surface Science **258**(16), 6080–6086 (2012)

[17] Block, S.B., da Silva, R.D., Dorini, L.B., Minetto, R.: Inspection of imprint defects in stamped metal surfaces using deep learning and tracking. IEEE Transactions on Industrial Electronics **68**(5), 4498–4507 (2020)

[18] Yang, Y., Yang, R., Pan, L., Ma, J., Zhu, Y., Diao, T., Zhang, L.: A lightweight deep learning algorithm for inspection of laser welding defects on safety vent of power battery. Computers in Industry **123**, 103306 (2020)

[19] Cha, Y.-J., Choi, W., Suh, G., Mahmoudkhani, S., Büyüköztürk, O.: Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. Computer-Aided Civil and Infrastructure Engineering **33**(9), 731–747 (2018)

[20] Li, J., Su, Z., Geng, J., Yin, Y.: Real-time detection of steel strip surface defects based on improved yolo detection network. IFAC-PapersOnLine **51**(21), 76–81 (2018)

[21] Zhuxi, M., Li, Y., Huang, M., Huang, Q., Cheng, J., Tang, S.: A lightweight detector based on attention mechanism for aluminum strip surface defect detection. Computers in Industry **136**, 103585 (2022)

[22] Zhang, J., Qian, S., Tan, C.: Automated bridge surface crack detection and segmentation using computer vision-based deep learning model. Engineering Applications of Artificial Intelligence **115**, 105225 (2022)

[23] Yao, J., Li, J.: Ayolov3-tiny: An improved convolutional neural network architecture for real-time defect detection of pad light guide plates. Computers in Industry **136**, 103588 (2022)

[24] Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)

[25] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[26] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

[27] Gao, L., Zhang, J., Yang, C., Zhou, Y.: Cas-vswin transformer: A variant swin transformer for surface-defect detection. Computers in Industry **140**, 103689 (2022)

[28] Severstal: Severstal: Steel defect detection. Kaggle (2018)

[29] He, Y., Song, K., Meng, Q., Yan, Y.: An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE Transactions on Instrumentation and Measurement **69**(4), 1493–1504 (2019)

[30] Božič, J., Tabernik, D., Skočaj, D.: Mixed supervision for surface-defect detection: from weakly to fully supervised learning. Computers in Industry **129**, 103459 (2021)

[31] Weimer, D., Scholz-Reiter, B., Shpitalni, M.: Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. CIRP Annals **65**(1), 417–420 (2016)

[32] Tabernik, D., Šela, S., Skvarč, J., Skočaj, D.: Segmentation-based deep-learning approach for surface-defect detection. Journal of Intelligent Manufacturing **31**(3), 759–776 (2020)

[33] Banterle, F., Artusi, A., Debattista, K., Chalmers, A.: Advanced High Dynamic Range Imaging, pp. 45–93. CRC press, New York (2017). https://doi.org/10.1201/9781315119526

[34] Rana, A., Valenzise, G., Dufaux, F.: Evaluation of feature detection in hdr based imaging under changes in illumination conditions. In: 2015 IEEE International Symposium on Multimedia (ISM), pp. 289–294 (2015). IEEE

[35] Mukherjee, R., Bessa, M., Melo-Pinto, P., Chalmers, A.: Object detection under challenging lighting conditions using high dynamic range imagery. IEEE Access **9**, 77771–77783 (2021)

[36] Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: ACM SIGGRAPH 2008 Classes, pp. 1–10 (2008)

[37] Nakajima, K., Kikuuma, T., Hasuka, K.: Yawata technical report no. 284. Yawata, Japan, 678–90 (1971)

[38] Reinhard, E., Stark, M., Shirley, P., Ferwerda, J.: Photographic tone reproduction for digital images. In: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, pp. 267–276 (2002)

[39] Larson, G.W., Rushmeier, H., Piatko, C.: A visibility matching tone reproduction operator for high dynamic range scenes. IEEE Transactions on Visualization and Computer Graphics **3**(4), 291–306 (1997)

[40] Tzutalin: LabelImg. GitHub. Accessed: 17-01-2023 (2015). https://github.com/tzutalin/labelImg/

[41] Ridnik, T., Lawen, H., Noy, A., Ben Baruch, E., Sharir, G., Friedman, I.: Tresnet: High performance gpu-dedicated architecture. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1400–1409 (2021)

[42] Reinhard, E.: Parameter estimation for photographic tone reproduction. Journal of graphics tools **7**(1), 45–51 (2002)