2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Deformable Dilated Faster R-CNN for Universal Lesion Detection in CT Images*

Fabio Hellmann[1], Zhao Ren[2], Elisabeth André[1], Björn W. Schuller[2,3]

***Abstract*— Cancer is a major public health issue and takes the second-highest toll of deaths caused by non-communicable diseases worldwide. Automatically detecting lesions at an early stage is essential to increase the chance of a cure. This study proposes a novel dilated Faster R-CNN with modulated deformable convolution and modulated deformable positive-sensitive region of interest pooling to detect lesions in computer tomography images. A pre-trained VGG-16 is transferred as the backbone of Faster R-CNN, followed by a region proposal network and a region of interest pooling layer to achieve lesion detection. The modulated deformable convolutional layers are employed to learn deformable convolutional filters, while the modulated deformable positive-sensitive region of interest pooling provides an enhanced feature extraction on the feature maps. Moreover, dilated convolutions are combined with the modulated deformable convolutions to fine-tune the VGG-16 model with multi-scale receptive fields. In the experiments evaluated on the DeepLesion dataset, the modulated deformable positive-sensitive region of interest pooling model achieves the highest sensitivity score of $58.8\,\%$ on average with dilation of $[4, 4, 4]$ and outperforms state-of-the-art models in the range of $[2, 8]$ average false positives per image. This research demonstrates the suitability of dilation modifications and the possibility of enhancing the performance using a modulated deformable positive-sensitive region of interest pooling layer for universal lesion detectors.**

I. Introduction

Cancer is a major public health issue worldwide. The World Health Organization (WHO) stated in their report [18] that cancer claimed 9.0 million among the overall amount of deaths due to Non-Communicable Diseases (NCDs), second to cardiovascular diseases only (17.9 million deaths). On average, a clinical diagnosis is provided after 4.6 minutes [7] by a physician to the patient, which limits the time for precise lesion detection and can cause up to 30 % of False Negative (FN) errors when scanning for lesions [14]. In this regard, a research study has shown that automatic lesion detection systems, i. e., Computer-Aided Detection (CADe), can detect up to 70 % of lung cancers which radiologists did not detect but missed 20 % of lung cancers identified by radiologists, which indicates the crucial assistant role of automatic lesion detection systems for physicians [16].

Recently, Universal Lesion Detection (ULD), as an essential technology of CADe systems, has shown effectiveness in many research studies [25], [2], [26], [22], [30]. The ULD focuses on detecting lesions throughout the whole body. The unique lesion shapes and the small number of training data make it very challenging to sufficiently train a model compared to other non-medical object detection tasks. However, traditional square convolutional kernels may become a bottleneck of achieving ULD with high performance when training with uniquely shaped lesions. Additionally, using a fixed size of convolutional kernels might affect the size of receptive fields, limiting to improve the performance while detecting specific sizes of lesions.

To overcome the above two limitations, we assume that unique sizes and shapes of lesions can be better recognized when adopting a universal lesion detector with dilation, Modulated Deformable Convolution (MDC), and Modulated Deformable Positive-Sensitive Region of Interest Pooling (MDPSRoIP). Since the increased dilation rate of convolutional layers enhances the size of receptive fields [12], the increased receptive field might help get additional information on the lesions' surroundings to classify them more accurately. Our previous studies [20], [21] in acoustic scene classification showed the dilated convolution's effectiveness in retaining high-resolution feature maps by increasing the receptive field's size. We assume that big-sized feature maps outputted by the backbone could be supportive in ULD as well. Moreover, deformable convolutional networks can tackle the lesions' unique shapes with two convolutional layers [3]. Specifically, one of the two convolutional layers provides an offset for the other layer's convolutional kernel. Further, another study [33] improved the deformable method to a modulated deformable method to perform better than the previous study [3]. More recently, lesion detection in mammography [15] successfully used the deformable convolution approach. Inspired by these studies [3], [33], [15], the MDC and MDPSRoIP are suspected to increase the layers' adaptability towards the unique shapes using offsets to transform the kernel reference point positions to a new location on the feature maps.

This work contributes three major improvements on the ULD performance of a typical Faster Region-Based Convolutional Neural Network (R-CNN) ULD on the DeepLesion dataset [27]. First, the layers in the last convolutional block of the VGG-16 backbone receive an increased dilation rate to cover more of the surrounding lesion areas and expect more accurate predictions. Further, MDCs replace the convolutional

*This work was supported by the European Union's Horizon H2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 766287 (TAPAS).

[1]Fabio Hellmann and Elisabeth André are with the Chair of Human - Centered AI, University of Augsburg, 86159 Augsburg, Germany {fabio.hellmann, andre}@informatik.uni-augsburg.de

[2]Zhao Ren and Björn W. Schuller are with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany {zhao.ren, schuller}@informatik.uni-augsburg.de

[3]Björn Schuller is also with the GLAM – the Group on Language, Audio, & Music, Imperial College London, SW7 2AZ London, UK

layers in the last convolutional block of the VGG-16 backbone of the Faster R-CNN for more adaptability to different lesion shapes. Finally, the Region of Interest (RoI) pooling layer is replaced with the MDPSRoIP layer to enhance the classification result.

II. Related Work

Lesion detection is a medical task of object detection. A recent study [13] described most architectures for object detection with a two-step process: i) a Region Proposal Network (RPN) processes the images with a backbone and uses the intermediate features to predict class-agnostic box proposals, and ii) a classifier predicts a class and class-specific box refinement for each proposal. Especially, the classifier differs for R-CNN and Region-Based Fully-Connected Network (R-FCN). R-CNN crops features from the same intermediate feature map as the RPN, while the R-FCN crops the features from the last layer before the prediction. A Multitask Universal Lesion Analysis Network (MULAN) [25] showed promising results on the DeepLesion [27] dataset for joint detection, tagging, and segmentation. MULAN used a Mask R-CNN framework with multiple head branches and a 3D feature fusion strategy [25]. However, MULAN struggled to learn well with the little amount of training data of the DeepLesion dataset [27]. Furthermore, another study [13] showed that Faster R-CNN is slightly better than the R-FCN and other architectures for detecting small objects. Therefore, in this paper, the Faster R-CNN is chosen due to the increased accuracy in detecting small objects.

Multi-Expert Lesion Detector (MELD) [23] tried to overcome the issue of a too-small amount of training data. Therefore, the study [23] introduced a framework to grasp the DeepLesion dataset and other datasets to train a multi-head multi-task lesion detector [23]. The trained MELD labeled those partially-labeled datasets, e. g., DeepLesion automatically[23]. To finetune MELD Missing Annotation Matching (MAM) and Negative Region Mining (NRM) were performed on training images to locate positive and negative areas [23]. The framework named Lesion ENSemble (LENS) [24] used a similar approach by combining the knowledge of several datasets, e. g., DeepLesion, but used an anchor-free proposal network instead of a RPN. However, our approach focuses only on the DeepLesion dataset to aim at higher sensitivity in detection with less data than that in MELD.

III. The Proposed Approach

This section introduces the structure of Faster R-CNN for ULD. Furthermore, the section presents the proposed dilations used in the Faster R-CNN and describes the modulated deformable operations in detail.

A. FASTER R-CNN

The Faster R-CNN embodies three components: backbone, RPN, and RoI pooling (cf. Figure 1). Each Computed Tomography (CT) image in the DeepLesion [27] dataset is firstly processed by the backbone (Conv1-5), which outputs a set of feature maps with a dimension of 64×64. Then, these feature maps are used as the input of both the RPN and RoI Pooling Layer. The RPN outputs the classification results (i. e., a probability score) and bounding box regressions (i. e., a set of four coordinates – $x_{bottom,left}$, $y_{bottom,left}$, $x_{top,right}$, $y_{top,right}$ – for each box). Further, these classification results and bounding box regressions are fed into the RoI Pooling Layer with the feature maps from the backbone. The output of the RoI Pooling Layer is 7×7 feature maps, which are finally provided as the Classifier's input to calculate the classifications and bounding box regressions for the detection results. In the following, the backbone, RPN, and RoI pooling will be introduced.

1) Backbone: The convolutional blocks (Conv1-5) inside a VGG-16 model are employed to extract abstract feature maps. For training time reduction, the backbone used a pre-trained VGG-16 model [4]. Particularly, while freezing the weights of the first two convolutional blocks (Conv1-2), only the high-level convolutional layers (Conv3-5) needed to be trained. Each convolutional block consists of convolutional layers with a Rectified Linear Unit (ReLU) and a max-pooling layer in the end. Due to the sparse number and often small lesions, the last two pooling layers (in Conv4-5) are removed (Pool4-5) [22]. Without the pooling layers, the feature map resolution remains the same throughout Conv4 to Conv5, allowing to detect lesions with small sizes and improve the positive sampling ratio (proposed regions that are True Positives (TPs)) [26]. At this point, the RPN and RoI Pooling Network receive the extracted feature maps from the VGG-16 backbone (Conv1-5).

2) Region Proposal Network: The specialized architecture of the RPN generates proposals for regions where objects could lie and their respective score of certainty. Anchors are in the central point of the sliding window with a variety of shapes. The anchor defines its width and height by the scales s and the aspect-ratios r. The number of possible proposals k for each pixel depends on the number of scales and aspect-ratios. For example, the anchor scale $s = [16]$ and aspect-ratios $r = [1:2, 1:1, 2:1]$ results in $A = 3$ anchor windows of the size of 16×32, 16×16, and 32×16. The whole image can have a maximum number of anchors to $W * H * A$, where W is the image's width, and H is the image's height. To produce anchor coordinates and a classification score, the RPN uses a convolutional layer as an intermediate layer to process the provided feature maps consumed by two sibling convolutional layers – a Box-Regression Layer (BRegL) and a Box-Classification Layer (BClsL) [19]. The BClsL result has the same dimension as the number of anchors A are processed. The BRegL outputs four times the dimension of the number of anchors A. In the final processing step, both outputs (BRegLs and BClsLs) transform into a tensor of $g \times 4$ for regression and $g \times 1$ for classification results, where g is the maximum number of anchors. Non-maximum Suppression (NMS) reduces the number of proposals and speeds up the training process with the regression results, classification results, and a threshold t_{nms}. The NMS removes lower-scoring boxes during its iterative operation, which have

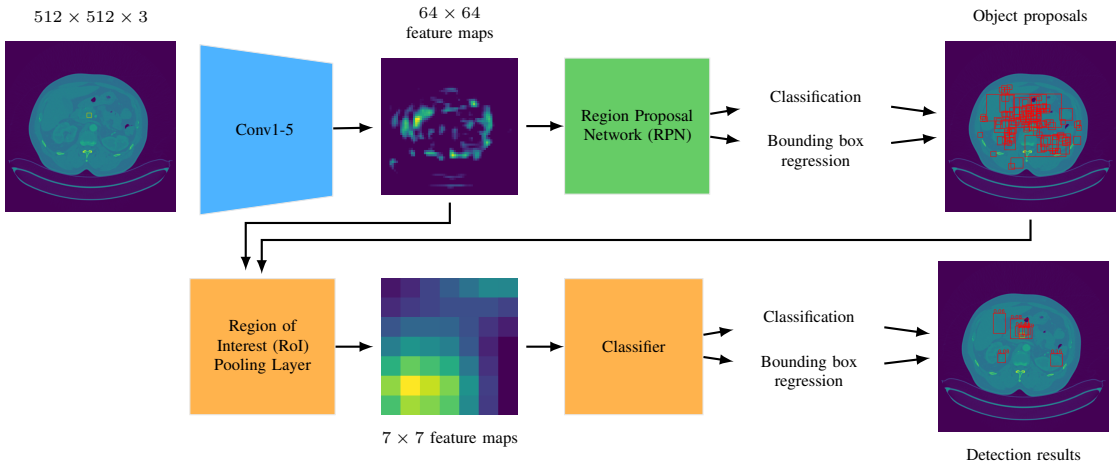

Fig. 1. An overview of the Faster R-CNN structure for ULD in CT images adopted from [11].

an Intersection-over-Unit (IoU) greater than t_{nms} with another higher-scoring box. The *overlap* divided by the *union area* results in the IoU, which is a percentage output.

3) Region of Interest Pooling: Faster R-CNN introduced the RoI Pooling Layer to perform max-pooling on inputs of unequal sizes to generate equal-size feature maps [9]. Therefore, the region proposal marks the area on the feature map where the RoI pooling performs [10]. This area is split into several bins according to the pool size (e. g., a pool of 2×2 has four bins) [10]. A max-pooling is performed for each of these bins to gain the bin area's maximum value [10]. These separations into bins lead to the known issue of misalignments between the RoI and the extracted features [10]. This paper overcomes these misalignments by using the RoI alignment [10], [31]. After the RoI processing step, the shrunk feature maps are forwarded to the classifier, which embodies the classifier of VGG-16 and two siblings Fully-Connected Layers (FCLs) – a BRegL and BClsL. In our model, convolutional layers (Conv6 and Conv7) replaced the FCLs in the VGG-16 classifier. According to a study [26], the replacement of the FCL with convolutional layers cuts the size of the model to $\frac{1}{4}$ while the accuracy remains nearly the same. However, Conv6 embodies 512-dimensions and has a kernel size of 3×3 and Conv7 consists of 512-dimensions and has a kernel size of 5×5 – both with zero padding and stride one. After processing the input through the VGG-16 classifier, the BRegL and BClsL process the output. For prediction purposes only, NMS is used to eliminate overlapping region detections with an IoU threshold of 0.5 but a minimum score of 0.05. These settings filter out all region detections with a score lower than 0.05, which is most certainly no lesion, and with an IoU higher than 0.5, because it would overlap with another region more than $50\,\%$, which would cover the lesion area as well. The softmax function computes the prediction score.

4) Training: Recent research [26] showed that end-to-end joint training could achieve the best results. The complete model embodies four different losses in the tasks of box-classification, and box-regression for each of RPN and R-CNN. During training, the RPN takes all anchors and puts those into the "foreground" category, which have an IoU of 0.5 or greater with a ground-truth object. The "background" category contains the anchors which do not overlap or have an IoU less than 0.1 with a ground-truth object. A mini-batch embodies 32 randomly sampled foreground and background anchors with a balanced ratio. The RPN calculates the classification loss with a Binary Cross-Entropy (BCE) [28] loss function based on the anchors contained in the mini-batch. The R-CNN uses a similar approach with the anchors as the RPN. The anchors are labeled as foreground and background again. However, those anchors, without any intersection, are ignored to focus on the more promising anchors. The randomly sampled balanced mini-batch of size 32 contains $25\,\%$ foreground and $75\,\%$ background proposals. The Categorical Cross-Entropy (CCE) [32] function computes the classification loss. Finally, the box-regression for both, RPN and R-CNN, uses the smooth L1 loss, as suggested in [9].

B. Dilation

Dilated convolution (or atrous convolution) increases the receptive field [12]. When using a dilation factor f with a kernel size s, the receptive field increases respectively. The size of the *effective kernel* $\hat{s}$ measures by $\hat{s} = s + (s-1)(f-1)$ [5]. Changing the dilation of a convolutional operation does not affect the number of parameters [29]. In our model, we take advantage of the increased receptive field when using dilation to cover more of the lesion's area and the surroundings. Therefore, the convolutional layers in the Conv5 block use dilation variations, as a study [8] stated that using dilation in the Conv5 block enhances the accuracy most.

C. Modulated Deformable Convolution

The 2D convolution processes a sample over the input feature map and sums up the sampled values weighted by iterating through the kernel locations. An additional convolutional layer provides the original convolution's kernel's sampling locations with 2D offsets. The offset allows

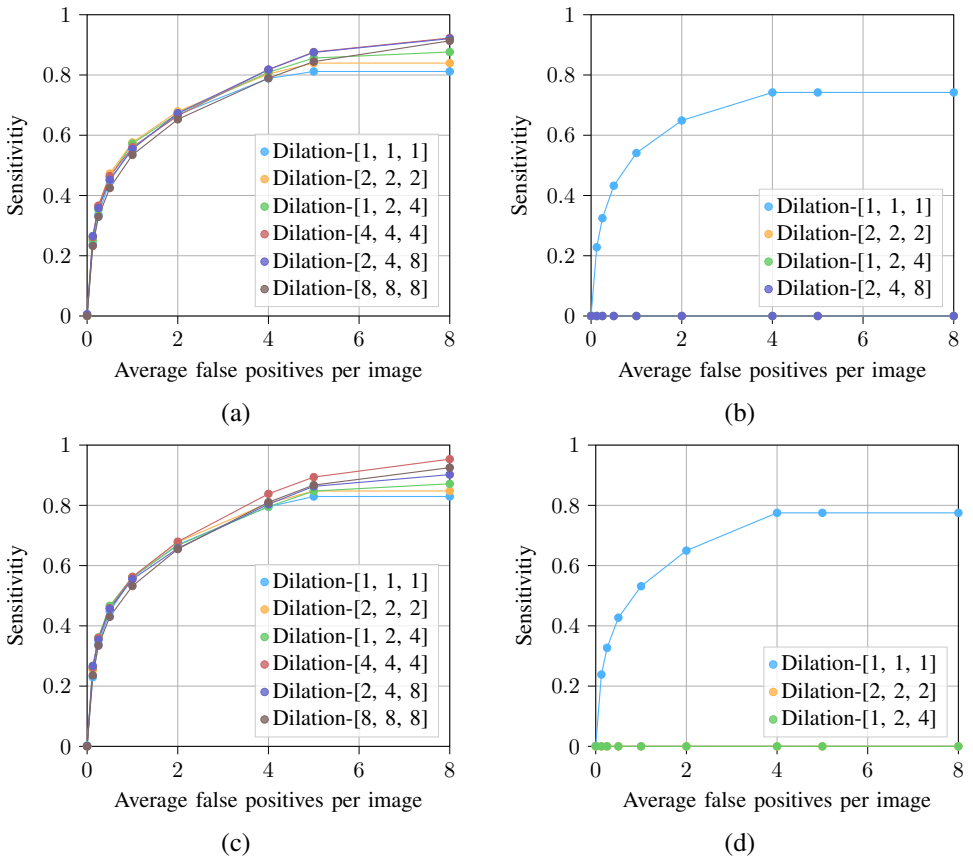

Fig. 2. The depicted charts represent the FROC curves of (a) a non-deformable model, (b) a MDC model, (c) a MDPSRoIP model, and (d) a combined MDC and MDPSRoIP model adopted from [11].

the sampling grid free form deformations [3]. A follow-up study [33] improved the deformable convolutional method by enhancing the focus on pertinent image regions. Therefore, the deformable convolution uses a modulation scalar on each kernel bin location. It is firstly proposed in this paper to use MDCs for ULD as a replacement for the convolutional layers in the Conv5 block. Additionally, the MDCs are used with dilation variations, the same as the standard convolutional layers, to evaluate if it enhances the accuracy.

D. Modulated Deformable Positive-Sensitive Region of Interest Pooling

Region proposal-based object detection methods use the RoI pooling operation as it transforms irregularly sized rectangular regions into fixed-sized features [3]. The standard RoI pooling uses the input feature map and RoIs. The deformable RoI pooling [3] uses offsets added to the spatial binning locations. Compared to the deformable RoI pooling method, the approach of the deformable Positive-Sensitive (PS) RoI pooling [3] is fully convolutional. A follow-up study [33] adapted the deformable RoI pooling by adding a modulation scalar to each bin which slightly increased the overall detection accuracy. This paper firstly proposes to use the MDPSRoIP as an alternative to RoI alignment for ULD with Faster R-CNN. The use of MDPSRoIP enhances the prediction results due to increased flexibility of the kernel.

IV. Experiments

The conducted experiments use the DeepLesion [27] dataset with the official split, which embodies in total 32 735 lesions in 32 120 CT slices from 10 594 studies of 4 427 patients. The dataset contains various lesion types, such as lung nodules and liver tumors. The DeepLesion dataset provides the CT slices as 12-bit images and their corresponding lesion annotations. To enhance the visibility of the tissue to analyze, such as lung, soft tissue, bones, the intensity range of the image x_{in} is transformed to a floating-point number after subtracting 32 768 using

$$x_{out}(x_{in}) = [(x_{in} - 32768) - u_{min}]/(u_{max} - u_{min}). \quad (1)$$

A single window of Hounsfield Unit (HU) was used to range it between $u_{min} = -1024$ and $u_{max} = 3071$ [26]. Afterward, all images were normalized and resized to 512 pixels in width and height along with the annotations' respective size and location. For gathering 3-dimensional information, three input slices were stacked together by using the annotated slice in the middle and the neighboring slices above and below. In case that the slice with the annotation is the first or last of all CT slices taken for this patient, the same slice was used twice to fill the empty remaining slice stack. The Free-response Receiver Operating Characteristic (FROC) curve [17] was used to evaluate all models' experiments' results. The models' training took between 1.5 to 3.5 hours/epoch on an NVIDIA

GeForce GTX 1080 Ti GPU. The source code of our work is released in GitHub[1].

A. Settings

Every experiment used the stochastic gradient descent [1] optimizer with a learning rate of 0.002, which was reduced after six epochs by a factor of 10. The training procedure took eight epochs, as a study [26] stated that the network would converge within this time. Furthermore, the model is fed with mini-batches of 8 images each. The losses of RPN and R-CNN were jointly optimized to be more efficient than optimizing them separately [19]. The RPN uses anchor scales of 16, 24, 32, 48, and 96 and ratios of 1:2, 1:1, and 2:1 as it was claimed best by a study [26]. The number of candidate lesion regions by the RPN is limited to 32 to reduce training time. The individual adoptions for each experiment embody the combination of a set of dilations with corresponding paddings for the last block (Conv5) in the backbone and the use of normal convolution/RoI pooling and MDC/MDPSRoIP. Replacing the last convolutional layers of Conv3-5 blocks with MDC layers within the backbone was tested with unsatisfactory results. However, the dilation and padding combinations used for the experiments are $[1, 1, 1]$, $[2, 2, 2]$, $[4, 4, 4]$, $[8, 8, 8]$, $[1, 2, 4]$, and $[2, 4, 8]$, where the number in brackets represents the dilation and padding for each of the three convolutional layers, respectively. The adjusted padding must keep the generated feature maps' size simultaneously throughout the different models. In total, with all variations, 24 models were trained during the experiment.

B. Results

The FROC curves in Figure 2 present the results of each model. The models with a MDC layer achieve their highest sensitivity score on average with dilation of $[1, 1, 1]$. In contrast, the other dilation variations stagnate with zero sensitivity due to a respectively high RPN class loss after eight epochs trained or result in a not-a-number/infinite value during the first epoch of training. The RPN class loss stagnation was confirmed in a test with a MDC model trained for 17 epochs without using a learning rate decay. However, the non-deformable and MDPSRoIP models gain a sensitivity value for all dilation variations. Both model types achieve their highest sensitivity score on average with a $[4, 4, 4]$ dilation. The overall highest sensitivity score on average of all models in this experiment achieved the MDPSRoIP model with dilation of $[4, 4, 4]$ and a sensitivity score on average of $58.8\,\%$. The second best is the non-deformable model with dilation of $[4, 4, 4]$ with a sensitivity score on average of $58.0\,\%$. Both models could reach the highest sensitivity scores for the average false positives per image greater than False Positive (FP)@2. Furthermore, both models hang slightly behind their counterparts, with dilation of $[2, 2, 2]$ or $[1, 2, 4]$, in the midsection between FP@0.25 and FP@2. Additionally, the models with dilation of $[2, 4, 8]$ reach the highest sensitivity score of the lowest FP@0.125 while their counterparts, with dilation of $[4, 4, 4]$, stay slightly behind $(0.3-0.5\,\%)$. The "FP@n" refers to the average false positives per image value n on the x-axis of the FROC curve.

[1]https://github.com/EIHW/Deformable_Dilated_Faster-RCNN

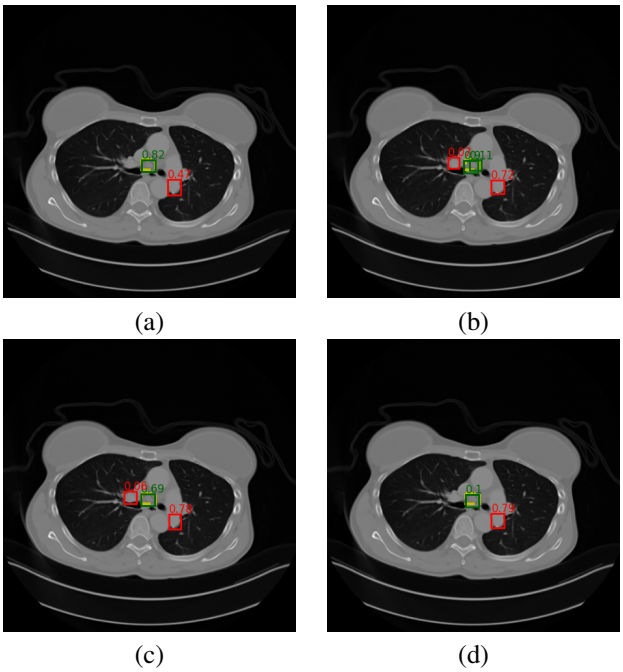

Fig. 3. A sample image processed by each model type with dilation of $[1, 1, 1]$: (a) non-deformable model, (b) MDC model, (c) MDPSRoIP model, and (d) combined MDC and MDPSRoIP model [11].

Figure 3 shows an exemplary output of the different models with a dilation of $[1, 1, 1]$. The non-deformable model detects the annotated lesions with a certainty of $82\,\%$ as well as one FPs with $47\,\%$. The MDC model on the other hand predicted with a much higher certainty of $90\,\%$ and a lower one with $11\,\%$ the TP lesion but increase the certainty for the FP lesion $72\,\%$ as well and even detected another FP lesion with $7\,\%$ certainty. The MDPSRoIP model achieved a certainty of $69\,\%$ for the TP lesion and predicts two FP lesions as well with $78\,\%$ and $6\,\%$ respectively. The combined deformable model has a distinct drop in prediction certainty over the TP lesion with $10\,\%$ but a higher certainty for the FP lesion than all the other models with $79\,\%$.

C. Discussion

The experiments in this paper evaluate the possible improvements of Faster R-CNNs for ULD. These possible enhancements cover the use of dilation, MDCs, MDPSRoIP layers, or a combination of them.

The FROC curves (Figure 2) show that the use of dilation can enhance the sensitivity for some of the average FPs per image. These findings can only be confirmed for the non-deformable model and MDPSRoIP model. The maximal enhanced results for the non-deformable and MDPSRoIP model with dilation of $[4, 4, 4]$ can be explained by the increased receptive field accordingly, additional information provided about the lesions. Compared to the other dilation

TABLE I

COMPARISON OF THE PERFORMANCE [%] STATE-OF-THE-ART METHODS WITH THE PROPOSED UNIVERSAL LESION DETECTION MODELS.

Method	Dilation	FP@ 0.125	0.25	0.5	1	2	4	8	Mean
MULAN [25], [24]		11.2	16.3	24.3	32.8	41.6	50.9	60.1	33.9
LENS [24]		23.7	31.6	40.3	50.0	59.6	69.5	78.0	50.4
MELD with MAM and NRM [23]		16.0	22.8	32.0	41.7	51.3	60.3	68.3	41.8
Ours									
Non-deformable model	$[4, 4, 4]$	**26.2**	**36.6**	**46.4**	56.0	66.4	81.8	92.3	58.0
Modulated deformable convolutional model	$[1, 1, 1]$	22.8	32.5	43.2	54.1	64.9	74.2	74.2	52.3
Modulated deformable PS ROI pooling model	$[4, 4, 4]$	**26.2**	36.0	45.8	**56.3**	**67.9**	**83.8**	**95.3**	**58.8**
Modulated deformable (conv. + PS ROI pooling) model	$[1, 1, 1]$	23.8	32.7	42.7	53.1	65.0	77.5	77.5	53.2

variations, the dilation with $[4, 4, 4]$ stands out in particular for FP@$[2, 4]$. Even though the sensitivity scores for FP@$[0.125, 1]$ lack slightly behind their counterpart models, the difference between them is not considerably different. Furthermore, the graphic shows that the models using MDC layers seem to adapt no longer when using different dilations than $[1, 1, 1]$ as the results remain at zero sensitivity. This sensitivity stagnation can be explained by the RPN class loss, which does not decrease considerably after the 5-6 epochs and accordingly generates random lesion proposals in the wrong places. The RPN class loss stagnation was confirmed in a separated test with a MDC model with dilation $[1, 2, 4]$ which trained for 17 epochs without a learning rate decay. Some of the MDC models with higher dilations (e. g., $[4, 4, 4]$, $[8, 8, 8]$) could not even finish the first epoch of training as the training interrupted with an error (not-a-number/infinite value), which is probably due to the increased number of zeros added by the padding and the behavior of deforming the kernel with an offset in combination.

The MDC model results contradict the improved results of deformable versus non-deformable convolutional layers as claimed in the study by Dai et al. [3]. The MDC model lacks high sensitivity scores compared to the non-deformable convolutional models. The missing pre-trained weights for the MDC layers can explain the lack of high sensitivity scores for the MDC model with dilation $[1, 1, 1]$ when comparing the non-deformable and modulated deformable backbone. The MDC layers replace the convolutional layers in the backbone after the pre-trained model was loaded. Accordingly, the MDC layers start with random weights, whereas the convolutional layers have pre-trained weights. Therefore, the MDC model needs more training epochs than the non-deformable convolutional models to reach the same or maybe higher sensitivity scores.

According to these findings, our models, with their dilation variations, are used to compare the results with other state-of-the-art studies (Table I). The columns within the "FP@" column represent the x-axis of the FROC curve. The mean column presents the mean over the values in range FP@$[0.125, 8]$. The highest sensitivity values are highlighted with bold characters for a better overview of the highest sensitivity scores. Comparing the values shows that the proposed models perform better in all segments compared to their state-of-the-art competitors. On average, the highest overall sensitivity score achieved our MDPSRoIP model with dilation of $[4, 4, 4]$. In a one-tailed z-test, the MDPSRoIP model scored with a significant improvement of $p < 0.001$ over the models in Table I (i.e., MULAN, LENS, and MELD with MAM and NRM) and the baseline non-deformable model with dilation of $[1, 1, 1]$ (55.0 %).

The results in Table I show high sensitivity scores in the FP@0.125 section with dilation of $[4, 4, 4]$ and $[2, 4, 8]$ compared to the other approaches. This fact is exciting in medical terms, as the lower FP range is critical in medical diagnosis due to the lower risk of false predictions. This finding indicates that using a MDPSRoIP approach can enhance the overall sensitivity compared to the non-deformable approach. It cannot be concluded if the MDPSRoIP performs better in general than the MDC model as the number of training epochs differs and was not evaluated. However, in terms of detected lesions, it should be kept in mind that in the DeepLesion dataset, those lesions marked as FPs may not be FPs because radiologists mark only relevant lesions [6]. Accordingly, by reducing the number of missed lesion annotations, the sensitivity scores could rise.

Another study [26] pointed out that rare lesion types, such as lung scarring, are often undetected in the DeepLesion dataset due to missing annotations from radiologists. Accordingly, the limitations, such as lack of complete labels in the dataset and noise in lesion annotations, from the related study [26] remain in this paper as well. Additionally, the present paper's results (Table I) are based on a 2.5-dimensional method and are not transferable to a 3-dimensional approach, as the results might differ.

V. CONCLUSION AND OUTLOOK

This research aimed to evaluate the suitability of dilation in non-deformable/Modulated Deformable Convolutions (MDCs) and the use of Modulated Deformable Positive-Sensitive Region of Interest Pooling (MDPSRoIP) compared to Region of Interest (RoI) alignment in Universal Lesion Detection (ULD). The last convolutional block of the VGG-16 backbone was modified by altering the dilation and padding, respectively, and replacing normal convolutional layers with MDCs. The RoI alignment component in the Faster Region-Based Convolutional Neural Network (R-CNN) was used and replaced with their MDPSRoIP counterpart. The conducted experiments applied the Free-response Receiver Operating

Characteristic (FROC) curve for evaluation. As expected, the sensitivity scores were enhanced when using greater dilation than $[1, 1, 1]$, which is due to the larger receptive field that can gather more information about a lesion area. Furthermore, the MDC was expected to enhance the sensitivity but instead reduced it due to the lack of pre-trained weights for the MDCs. Besides, dilation with MDC under-performed exceptionally, as the Region Proposal Network (RPN) class loss did not decline sufficiently fast in 8 or even 17 epochs to gain a good prediction result. However, the performance of the MDPSRoIP model presented an increase in sensitivity scores, especially in False Positive (FP)@$[2, 8]$ as well as the lower FP@0.125 rate, which is in particular important in medical terms to keep the false predictions at a minimum to reduce misdiagnosing patients. The MDPSRoIP model achieved the highest sensitivity on average with $58.8\,\%$ compared to other models presented in this research and other state-of-the-art methods.

In future efforts, the uncertainty of the models should be thoroughly evaluated. Moreover, the learning rate stagnation of MDC models with an increased dilation needs further investigation to evaluate the true potential of MDC in ULD. Furthermore, the chance of detecting a lesion might be correlated to the size and location of the lesion as sensitivity scores fluctuate drastically [26]. Finally, 3D input processing needs evaluation as other research [2] claimed improved results with 3D computation.

REFERENCES

[1] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proc. COMPSTAT*, pages 177–186. Paris, France, 2010.

[2] J. Cai, K. Yan, C.-T. Cheng, J. Xiao, C.-H. Liao, L. Lu, and A. P. Harrison. Deep volumetric universal lesion detection using light-weight pseudo 3D convolution and surface point regression. In *Proc. MICCAI*, pages 3–13, 2020.

[3] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *ICCV*, pages 2380–7504, Venice, Italy, 2017.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, Miami, FL, 2009.

[5] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning, 2018. 31 pages.

[6] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised RECIST guideline. *European Journal of Cancer*, 45(2):228–247, Jan. 2009.

[7] S. Gaur et al. Can computer-aided diagnosis assist in the identification of prostate cancer on prostate MRI? A multi-center, multi-reader investigation. *Oncotarget*, 9(73):33804–33817, Sep. 2018.

[8] L. Geng, S. Zhang, J. Tong, and Z. Xiao. Lung segmentation method with dilated convolution based on VGG-16 network. *Computer Assisted Surgery*, 24(sup2):27–33, Aug. 2019.

[9] R. Girshick. Fast R-CNN. In *Proc. ICCV*, Santiago, Chile, 2015. 9 pages.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, June 2020.

[11] F. Hellmann. Deformable Faster-RCNN for lesion detection in CT images. Master's thesis, University of Augsburg, Augsburg, Germany, 2020. 51 pages.

[12] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Proc. TFMPS*, pages 286–297, Marseille, France, 1990.

[13] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proc. CVPR*, pages 7310–7319, Honolulu, HI, 2017.

[14] H. L. Kundel, C. F. Nodine, and D. Carmody. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology*, 13(3):175–181, May 1978.

[15] C. Li, D. Zhang, Z. Tian, S. Du, and Y. Qu. Few-shot learning with deformable convolution for multiscale lesion detection in mammography. *Medical Physics*, 47(7):2970–2985, July 2020.

[16] M. Liang, W. Tang, D. M. Xu, A. C. Jirapatnakul, A. P. Reeves, C. I. Henschke, and D. Yankelevitz. Low-dose CT screening for lung cancer: Computer-aided detection of missed lung cancers. *Radiology*, 281(1):279–288, Mar. 2016.

[17] H. Miller. The FROC curve: A representation of the observer's performance for the method of free response. *The Journal of the Acoustical Society of America*, 46(6B):1473–1476, Dec. 1969.

[18] W. H. Organization. *World health statistics 2020: Monitoring health for the SDGs, sustainable development goals*. 2020.

[19] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2016.

[20] Z. Ren, Q. Kong, J. Han, M. Plumbley, and B. Schuller. Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes. In *Proc. ICASSP*, pages 56–60, Brighton, UK, 2019.

[21] Z. Ren, Q. Kong, J. Han, M. Plumbley, and B. Schuller. CAA-Net: Conditional atrous CNNs with attention for explainable device-robust acoustic scene classification. *IEEE Transactions on Multimedia*, Nov. 2020. 12 pages.

[22] K. Yan, M. Bagheri, and R. M. Summers. 3D context enhanced region-based convolutional neural network for end-to-end lesion detection. In *Proc. MICCAI*, pages 511–519, Granada, Spain, 2018.

[23] K. Yan, J. Cai, A. P. Harrison, D. Jin, J. Xiao, and L. Lu. Universal lesion detection by learning from multiple heterogeneously labeled datasets, 2020. 21 pages.

[24] K. Yan, J. Cai, Y. Zheng, A. P. Harrison, D. Jin, Y. Tang, Y. Tang, L. Huang, J. Xiao, and L. Lu. Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in CT. *IEEE Transactions on Medical Imaging*, Dec. 2020.

[25] K. Yan, Y. Tang, Y. Peng, V. Sandfort, M. Bagheri, Z. Lu, and R. M. Summers. MULAN: Multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In *Proc. MICCAI*, pages 194–202, Shenzhen, China, 2019.

[26] K. Yan, X. Wang, L. Lu, and R. M. Summers. DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3):036501–1–036501–11, July 2018.

[27] K. Yan, X. Wang, L. Lu, L. Zhang, A. P. Harrison, M. Bagheri, and R. M. Summers. Deep lesion graph in the wild: Relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In *Proc. CVPR*, pages 413–435, Salt Lake City, UT, 2019.

[28] M. Yi-de, L. Qing, and Q. Zhi-bai. Automated image segmentation using improved PCNN model based on cross-entropy. In *Proc. ISIMP*, pages 743–746, Hong Kong, China, 2004.

[29] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *Proc. ICLR*, San Juan, Puerto Rico, 2016. 13 pages.

[30] N. Zhang, Y. Cao, B. Liu, and Y. Luo. 3D aggregated faster r-cnn for general lesion detection, 2020. 11 pages.

[31] X. Zhang, K. Zhu, G. Chen, X. Tan, L. Zhang, F. Dai, P. Liao, and Y. Gong. Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network. *Remote Sensing*, 11(7), Mar. 2019. 27 pages.

[32] Z. Zhang and M. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proc. NIPS*, Red Hook, NY, 2018. 8792-–8802.

[33] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable ConvNets v2: More deformable, better results. In *Proc. CVPR*, pages 9308–9316, Long Beach, CA, 2019.