DKMA-ULD: Domain Knowledge augmented Multi-head Attention based Robust Universal Lesion Detection

Manu Sheoran*
manu.sheoran@tcs.com
Meghal Dani*
dani.meghal@tcs.com
Monika Sharma
monika.sharma1@tcs.com
Lovekesh Vig
lovekesh.vig@tcs.com

TCS Research
New Delhi, India

Abstract

Incorporating data-specific domain knowledge in deep networks explicitly can provide important cues beneficial for lesion detection and can mitigate the need for diverse heterogeneous datasets for learning robust detectors. In this paper, we exploit the domain information present in computed tomography (CT) scans and propose a robust universal lesion detection (ULD) network that can detect lesions across all organs of the body by training on a single dataset, DeepLesion. We analyze CT-slices of varying intensities, generated using heuristically determined Hounsfield Unit (HU) windows that individually highlight different organs and are given as inputs to the deep network. The features obtained from the multiple intensity images are fused using a novel convolution augmented multi-head self-attention module and subsequently, passed to a Region Proposal Network (RPN) for lesion detection. In addition, we observed that traditional anchor boxes used in RPN for natural images are not suitable for lesion sizes often found in medical images. Therefore, we propose to use lesion-specific anchor sizes and ratios in the RPN for improving the detection performance. We use self-supervision to initialize weights of our network on the DeepLesion dataset to further imbibe domain knowledge. Our proposed Domain Knowledge augmented Multi-head Attention based Universal Lesion Detection Network **DMKA-ULD** produces refined and precise bounding boxes around lesions across different organs. We evaluate the efficacy of our network on the publicly available DeepLesion dataset which comprises of approximately $32K$ CT scans with annotated lesions across all organs of the body. Results demonstrate that we outperform existing state-of-the-art methods achieving an overall sensitivity of 87.16%.

1 Introduction

Advances in deep learning techniques have led to significant breakthroughs in medical image analysis [15, 16, 19, 21]. In the past, efforts have been made to build automated lesion

* Equal contribution in paper

detection solutions that focus on specific organs such as liver, kidney, and lungs [12, 28, 36]. However, to address the clinical necessity where radiologists are required to locate different types of lesions present in various organs of the body to diagnose patients and determine treatment, developing a universal lesion detection (ULD) [13, 23, 32, 34, 38] model has become an active area of research. Tang et. al [23] proposed ULDor based on Mask-RCNN for lesion detection and a hard negative mining (HNEM) strategy to reduce false positives. However, the proposed HNEM technique may not enhance detection performance due to missing annotations as the mined negatives may actually contain positives. There are a few more impressive RCNN based ULD networks that use weights pre-trained on Imagenet for detection [18, 32, 34]. We also found from earlier methods [29, 32] that utilizing neighboring slice information is essential for providing 3D context to the network which gives a lift in lesion detection accuracy. This is due to the fact that clinicians look at multiple slices of a patient's CT scan to confirm the final diagnosis.

There also exist attention-based ULD networks where attention has been shown to improve the lesion detection [13, 24, 29, 37] by enabling the network to focus on important regions of CT-scans. MVP-Net [13] proposed to use a position-aware attention module to aggregate features from a multi-view feature pyramid network. Another work on ULD by Wang et al. [29], proposed volumetric attention which exploits 3D-context from multi-slice image inputs and a 2.5D network for improving the detection performance. The multi-task universal lesion analysis network (MULAN) [34] utilizes 27 slices as input and proposes a 3D feature fusion strategy with Mask-RCNN backbone for lesion detection. In addition, they jointly train the network to perform lesion segmentation and tagging.

Typically, deep networks are reliant on high-volume datasets for automatically discovering relevant features for a learning task. However, due to the very similar appearance of lesions and other internal structures in CT scans, lesion detection is quite a challenging problem. Yan et al. [31] proposed a Lesion ENSemble (LENS) network for lesion detection that can efficiently learn from heterogeneous lesion datasets and address the issue of missing annotations by exploiting clinical prior knowledge and cross-dataset knowledge transfer. In another paper [30], authors have proposed a MELD network for lesion detection which learns from multiple heterogeneous diverse datasets and uses missing annotation matching (MAM) and negative region mining (NRM) for achieving state-of-the-art lesion detection performance on DeepLesion [33] dataset. In summary, previous works on ULD have made use of 3D-context in the form of multi-slice inputs, incorporation of attention mechanisms, multi-task learning, hard negative mining techniques, and multiple heterogeneous datasets to enhance the lesion detection sensitivity performance.

Rather than using a variety of heterogeneous datasets to learn robust representations for ULD, we claim that significant improvements to learning can be obtained by incorporating task-specific domain knowledge in the network. This motivates us to extract as many domain-specific features as possible from a minimal number of CT-slices for a particular patient, to come up with a computationally efficient ULD with enhanced prediction performance. Therefore, in our proposed lesion detector, we utilize only 3 slices from a patient's CT scan to incorporate 3D context in the network. Next, taking cues from MVP-Net [13], we utilize the information of tissue density from CT-scans represented as HU-values in terms of window width and window length. During manual analysis, radiologists adjust these windows to focus on organs/tissues of interest [2]. Despite having critical importance in lesion detection, the use of multiple HU windows of CT-slices, representing different organs of the body, has been overlooked in the literature. To this end, we introduce 5 novel heuristically determined HU windows for CT-slices and feed them as multi-intensity input to the detec-

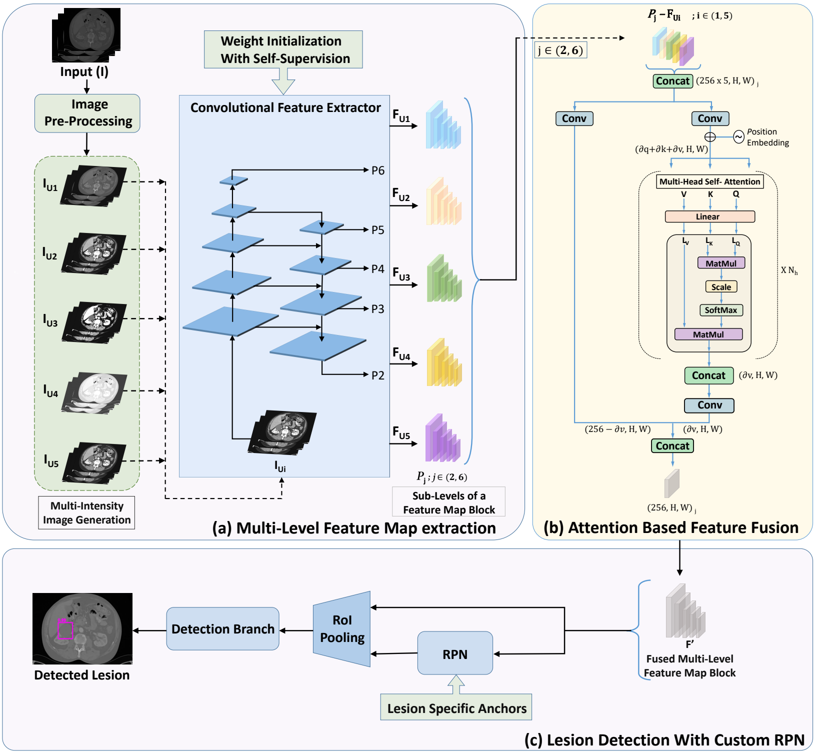


Figure 1: Overview of our *DKMA-ULD* architecture. (a) An input image I, consisting of 3 CT-slices of a patient, is pre-processed and used for generating 5 multi-intensity images using 5 different HU windows (U_1 to U_5). Using a shared convolutional feature extractor (having domain-specific weight initialization) with 5 FPN levels (P_2 to P_6), we obtain 5 feature map blocks F_{Ui}, (b) Using our proposed attention based feature fusion module inspired from transformer's multi-head self-attention [27], feature map sub-levels $(P_j - F_{Ui})$ with same resolution of (H,W) are fused into a single sub-level for obtaining the final fused feature map block (F′). Here V, K, Q and N_h represent value, key, query matrix and number of attention heads and finally, (c) RPN with custom lesion-specific anchors is applied over F′ for improved lesion detection.

tion network to make it organ-agnostic. Further, the computed features are combined using our novel convolution augmented multi-head attention-based fusion architecture. We use transformers [5, 7, 27] based self-attention mechanism for feature-fusion of multi-intensity images that effectively combines multi-organ information efficiently. This is analogous to the radiologists' way of paying attention to different organs at different HU windows of CT-slices simultaneously while detecting lesions. Additionally, we have observed that default anchor sizes and ratios used in general object detection networks do not perform satisfactorily for lesions of different sizes, particularly for very small ($<$ 10mm) and medium-sized ($10-30$mm) lesions. Therefore, we propose new anchor sizes and ratios for RPN that enable the network to detect lesions of varied sizes mostly found in medical imaging datasets. We named our network ***DKMA-ULD*** (Domain-knowledge augmented multi-attention based universal lesion detection).

Furthermore, observing that self-supervised learning (SSL) techniques have recently become the cornerstone for learning improved representations in data-scarce scenarios which can, subsequently, be used for efficient learning on downstream tasks [6, 9, 26]. We utilize Bootstrap Your Own Latent (BYOL)[9], a SSL technique to learn weights of the backbone network of *DKMA-ULD* on the DeepLesion [33] dataset. The DeepLesion dataset consists of approximately $32K$ CT scans with annotated lesions across different organs of the body. Subsequently, the weights learned via SSL are used for initializing our *DKMA-ULD* so that it is able to learn robust domain-specific representations resulting in improved sensitivity. To summarize, we make the following contributions in the paper:

- We propose a domain-knowledge augmented deep network with multi-head self-attention based feature-fusion named *DKMA-ULD* which performs robust detection of lesions across all organs of the body and is trained on DeepLesion [33] dataset.
- We introduce 5 novel HU windows, computed in an unsupervised manner, for highlighting the different organs of the body in CT -scans which make our network organ-agnostic.
- We propose a novel convolution augmented multi-head self-attention mechanism for the fusion of the features obtained from multiple intensity CT-slices for subsequent detection by an RPN.
- We propose lesion-specific new anchor sizes and ratios for detection that cover various sizes of lesions present in medical images. We also illustrate that these new anchors can detect very small and medium-sized lesions effectively. Hence, giving a boost to the overall detection sensitivity.
- We demonstrate that initializing DKMA-ULD with self-supervised domain-specific weights is beneficial for learning improved representations over Imagenet weights.
- We evaluate *DKMA-ULD* on DeepLesion dataset and show improvement over existing state-of-the-methods of ULD such as MULAN [34], 3DCE [32], MVP-Net [13], MELD [30] and improved RetinaNet [37].

2 Dataset Details

DeepLesion [33] is the largest publicly available repository of CT-slices with annotated lesions across different organs of the body released by the National Institutes of Health (NIH). It consists of data from $4,427$ unique patients based on markings performed by radiologists during their routine work. There are approximately $32,120$ axial CT slices from $10,594$ CT studies of the patients having around 1-3 lesions annotated per CT -scan. The lesions in each image have annotations such as bounding-box coordinates and size measurements, etc. which add up to $32,735$ lesions altogether from eight different body organs including bone, abdomen, mediastinum, liver, lung, kidney, soft-tissue, and pelvis. We use the official split of the DeepLesion dataset for training and evaluation of our proposed *DKMA-ULD*.

3 Proposed Method: *DKMA-ULD*

DKMA-ULD method, as shown in Figure 1, consists of different modules such as pre-processing, multi-intensity image generation using five different HU windows, convolutional feature extraction backbone, multi-head self-attention based feature-fusion, lesion-specific anchors, and self-supervision. These modules are discussed in detail as follows:

- **Pre-processing**: Typically, clinicians observe multiple adjacent slices of a patient's CT scan to confirm the diagnosis of a pathology. In order to provide 3D context of a patient's CT-scan to the network, we utilize its 3 slices (key slice with one superior

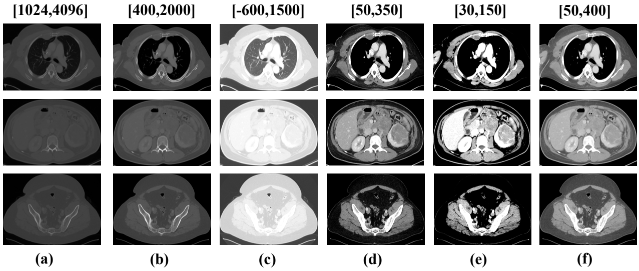


Figure 2: The top, middle, and bottom rows have CT-slices of chest-region, abdomen-region, and pelvic-region, respectively. The column (a) illustrates images with commonly used HU window ($U = [1024, 4096]$) while columns ranging from (b) to (f) have images with our new 5 HU windows. The figure clearly demonstrates that by using more number of windows, different organs of a particular body-region present in a CT volume, can be highlighted more efficiently.)

and one inferior neighboring slice) to generate 3-channel image. First, we remove black borders of the CT-slices for computational efficiency and to focus on region-of-interest. Next, we normalize and clip the 12-bit intensity values of a CT slice using a HU window and then, re-scale them to floating-point values in the range $[0.0, 255.0]$. Subsequently, we re-sample all the CT-slices to a common resolution of $0.8 \times 0.8 \times 2$ mm^3. In addition, since the DeepLesion dataset does not have segmentation masks for the lesions, we generate pseudo masks using provided RECIST diameter measurements [8]. These pseudo masks boost the performance of Mask-RCNN [10] by adding a branch for predicting segmentation masks on region proposals generated by RPN. We also augment the data during training using random affine transformations such as horizontal and vertical flips, resizing with a ratio of 0.8 to 1.2, and translation of $(-8, 8)$ pixels in x and y direction.

- **Multiple Intensity Image Generation**: In general, the intensity of a CT-slice is re-scaled using a certain HU window, U (e.g., a single and wide window of $[1024, 4096]$) in order to include gray-scale intensities of different organs [23, 32]. However, using a single window suppresses organ-specific information resulting into a degenerated image-contrast, as shown in Figure 2(a), which in turn makes it hard for the network to learn to focus on various organs present in the given CT volume. During manual detection of lesions, radiologists adjust these intensity values to focus on organs/tissue of interest [2]. We exploit this domain knowledge and propose to feed it to the deep network explicitly in the form of CT-slices having multiple intensities which highlight different organs of the body. In a previous method by Zihao Li et al. [13], a clustering algorithm is used to determine three HU windows. In this paper, we incorporate this multi-organ information in input CT-slices by introducing five novel HU windows which are determined in such a way that the major body organs are covered. The proposed HU windows, as inspired by Masoudi et al. [17], which cover almost all organs of interest for radiologists are: $U_1 = [400, 2000]$, $U_{2,3} = [-600, 1500], [50, 350]$, $U_4 = [30, 150]$, $U_5 = [50, 400]$ for bones, chest region including lungs & mediastinum, abdomen including liver & kidney, and soft-tissues, respectively. For a $U = [\mathbf{U_l}, \mathbf{U_w}]$, where, $\mathbf{U_l}$ and $\mathbf{U_w}$ are the window level/center and window width, the intensity values of a CT-slice are first normalized using $\mathbf{U_l} \pm \mathbf{U_w}/2$ as data min/max & clipped between $[0, 1]$ and then, re-scaled to values in $[0, 255]$.

- **Convolution Feature Extraction Backbone**: Now, for a given patient, 5 multiple intensity images each having 3 slices/channels, are passed as input to the ResNeXt-152 shared backbone with feature pyramid network (FPN) [14] based convolutional feature extractor. The fact that applying pooling layers of CNN repeatedly on an image can filter out information of small objects due to downsampling, hence, resulting in missing small and medium-sized lesions in radiological scans. Therefore, we utilize FPN where shallow and deeper feature maps are more suitable for detecting small/medium and larger lesions, respectively. As a result, for a given input, we obtain 5 feature-map blocks (F_{Ui}) corresponding to 5 FPN levels, each having 5 feature map sub-levels (P_j) of dimension $(256,H,W)_j$, where H and W represent height and width of the feature-map and $j = 2,...,6$ are the pyramid-levels. These extracted feature maps at different FPN levels, each having a different resolution allows RPN to effectively focus on lesions of different sizes.

- **Convolution augmented Multi-head Attention for Feature Fusion**: Earlier ULD techniques such as MULAN [34] incorporated information from multiple slices in their network by fusing the feature maps of all 3-channel images/slices with a convolution layer to obtain a 3D-context-enhanced feature map for the central slice. In this work, we propose a novel convolution augmented multi-head attention-based feature fusion module. Recently, vision transformers [6, 7, 25] have achieved state-of-the-art results on various machine vision tasks via a focus on self-attention [27]. The use of multi-head self-attention enables the model to attend jointly to both spatial and feature sub-spaces. As shown in Figure 1(b), we first concatenate sub-level feature maps $(256,H,W)_j$ of 5 different intensities to obtain feature-vectors of shape $(256*5,H,W)_j$. These features are, subsequently, passed to two parallel branches namely, the 2D convolution layer and transformer's multi-head self-attention [7]. Finally, their outputs are fused using concatenation. Since the output depth of the attention module is dependent on the depth of its "values" matrix (dv), the output depth of $2D$ convolution branch is kept such that the depth of the final feature vector obtained after concatenation of both the outputs is 256. Similar attention-based feature fusion is used at all 5 feature-map sub-levels and finally, we obtain a fused feature map block (F', with 5 feature-maps sub-levels) for later processing. To minimize computation overhead for attention, we use 2 attention heads (N_h) and keep the depth of values matrix as 4. In addition, we use 20 dimensions per head for key and query matrix [3].

- **Lesion-specific Anchors**: Next, for extracting Regions of Interest (ROI) from the obtained feature maps from 5 FPN levels, anchor boxes play a very crucial role. We observed that the small lesions are hard to detect using default anchor sizes and ratios [10, 14] used in RPN for real-world object detection. To circumvent this, we propose new custom anchors which are well suited for detecting lesions of all sizes in CT scans. Let's say, H and W are the image height and width, respectively. We generate anchor boxes of different sizes centered on each image pixel such that it has maximum IoU with lesion bounding box. If anchor sizes and ratios are in sets $\{s_1,s_2,...,s_n\}$ and $\{r_1,r_2,...,r_m\}$, respectively for each $r > 0$, we will have a total of $WH(n+m-1)$ anchor boxes [35]. Consider w_b and h_b to be the width and height of anchor boxes,

$$[w_b,h_b] = [Ws_n\sqrt{r_m}, Hs_n/\sqrt{r_m}] \quad s.t. \quad n,m \in [1,5] \tag{1}$$

We employ a differential evolution search algorithm [22] and find 5 best anchor sizes $[16,24,64,128,256]$ and ratios $[3.27,1.78,1,0.56,0.30]$ for $P2,P3,P4,P5,P6$ feature

Method	**Windows**	**FP@0.5**	**FP@1.0**	**FP@2.0**	**FP@4.0**	**Average**
3DCE(27 slices) [32]	1	62.48	73.37	80.70	85.65	75.55
improved RetinaNet(3 slices) [37]	1	72.18	80.07	86.40	90.77	82.36
MVP Net(9 slices) [13]	3	73.83	81.82	87.60	91.30	83.64
MULAN(27 slices, w/o tags) [34]	1	76.10	82.50	87.50	90.90	84.33
MULAN(27 slices, w/ 171 tags) [34]	1	76.12	83.69	88.76	92.30	85.22
MELD(9 slices)) [30]	1	77.80	84.80	89.00	91.80	85.90
MELD+MAM+NRM(9 slices) [30]	1	78.60	85.50	89.60	92.50	86.60
(a) *DKMA-ULD**	5	78.10	85.26	90.48	93.48	86.88
(b)**+Self-Supervision**	**5**	**78.75**	**85.95**	**90.48**	**93.48**	**87.16**

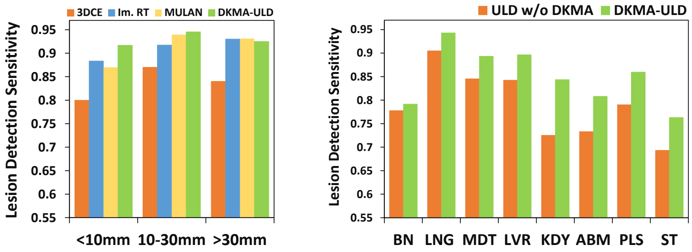
Table 1: Comparison of *DKMA-ULD* with previous state-of-the-art ULD methods. Sensitivity(%) at different false-positives (FP) per sub-volume on the volumetric test-set of DeepLesion [33] dataset. Here, we feed CT-slices, after performing cropping of black-borders during pre-processing, to the network. *training with Imagenet pre-trained weights.

map sub-levels, respectively. These lesion-specific anchors are used in RPN for RoI extraction, which are combined with feature maps using RoI pooling layer and further used, for predicting bounding-boxes around lesions along with probability values, as shown in Figure 1. We demonstrate that the custom anchors allow us to cover varied-sized lesions and more specifically, improve the detection of small-sized (< 10mm) and medium-sized ($10-30$mm) lesions considerably, as evident in Figure 3(a).

- **Self supervision**: The idea behind self-supervised learning (SSL) is that the learned intermediate representations can carry better semantic and structural meanings and can prove to be beneficial for a variety of downstream tasks. In order to make our *DKMA-ULD* more robust, we utilize a state-of-the-art SSL technique called BYOL [9]. It relies on two neural networks, referred to as online and target networks, that interact and learn from each other. The target network (parameterized by ξ) has the same architecture as the online one (parameterized by θ), but with polyak averaged weights, $\xi \leftarrow \tau\xi + (1-\tau)\theta$. The goal is to learn a representation y that can be used in downstream tasks. Generally, the detection networks are initialized using weights pre-trained on Imagenet consisting of natural images and may not be effective for the medical imaging domain. Therefore, we propose domain-specific weights, obtained by training the backbone using SSL over train-split (23K images) of the DeepLesion dataset, for initializing *DKMA-ULD* to obtain enhanced performance.

4 Experimental Results and Discussions

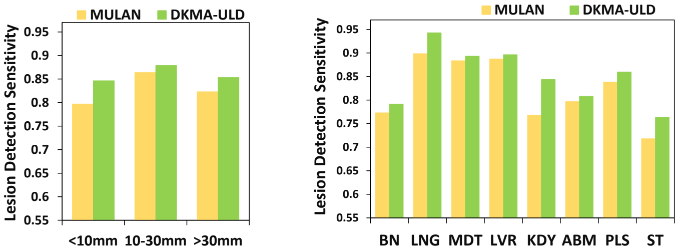
We use the official data split of DeepLesion [33] dataset which consists of 70%, 15%, 15% for training, validation, and test, respectively. Please note that the DeepLesion test-set includes only key CT-slices and may contain missing annotations. The lesion detection is classified as true positive (TP) when the IoU between the predicted and the ground-truth bounding-box is larger than 0.5. We report average sensitivity, computed at 0.5, 1, 2, and 4 false-positives (FP) per image, as the evaluation metric on the standard test-set split for the fair comparison.



a) Per Lesion Size Sensitivity Comparison

b) Per Organ Sensitivity Comparison

Figure 3: Sensitivity comparison for different lesion sizes and organs. a) Sensitivity (at FP= 4) for lesions with 3 different size ranges is compared with existing methods [32, 34, 37] and our proposed DKMA-ULD network w/o self-supervision. b) Average sensitivity comparison per organ computed using our proposed lesion detector without and with domain knowledge & multi-head attention. Here, BN, LNG, MDT, LVR, KDY, ABM, PLS and ST represent different organs such as bones, lungs, mediastinum, liver, kidney, abdomen, pelvis and soft-tissues, respectively.



a) Per Lesion Size Sensitivity Comparison

b) Per Organ Sensitivity Comparison

Figure 4: Average sensitivity comparison for different lesion sizes and organs. a) Average sensitivity ($FP = \{0.5, 1, 2, 4\}$) for lesions with 3 different size ranges is compared with MULAN [34] and our proposed *DKMA-ULD*. b) Organ-wise average sensitivity of MULAN and DKMA-ULD.

Training: The proposed *DKMA-ULD* is trained on 3 channel CT images of size 512×512 with a batch size of 4 on a single NVIDIA Tesla V100 having 32GB GPU-memory. We use cross-entropy and smooth ℓ_1 loss for classification and bounding-box regression, respectively. The model is trained until convergence using SGD optimizer with a learning rate (LR) and decay-factor of 0.02 and 10, respectively. The SSL model is trained using cross-entropy loss with a batch size of 64, Adam optimizer [11], and LR of $3e-4$ for 300 epochs.

Results and Discussions: We evaluate the performance of our *DKMA-ULD* against previous methods in the literature, as shown in Table 1. Our experiments demonstrate that by using only 3 slices per patient, the proposed method *DKMA-ULD* outperforms all the previous state-of-the-art ULD methods at different FP per image and achieves an average sensitivity of 86.88% when Imagenet pre-trained weights are used for the backbone initialization. Here, one important point to note is that the base model of MELD [30] (refer row 6 of Table 1) achieves an average sensitivity of 85.90% by training on 4 heterogeneous datasets, namely LUNA [20], LiTS [4], NIH-Lymph [1] and DeepLesion [33]. On the other hand, our proposed *DKMA-ULD* base model w/o self-supervision, despite being trained only on DeepLesion train-set, beats MELD on Deeplesion test-set by achieving an average sensitivity of 86.88%. Therefore, it is evident that the introduction of domain-specific features and

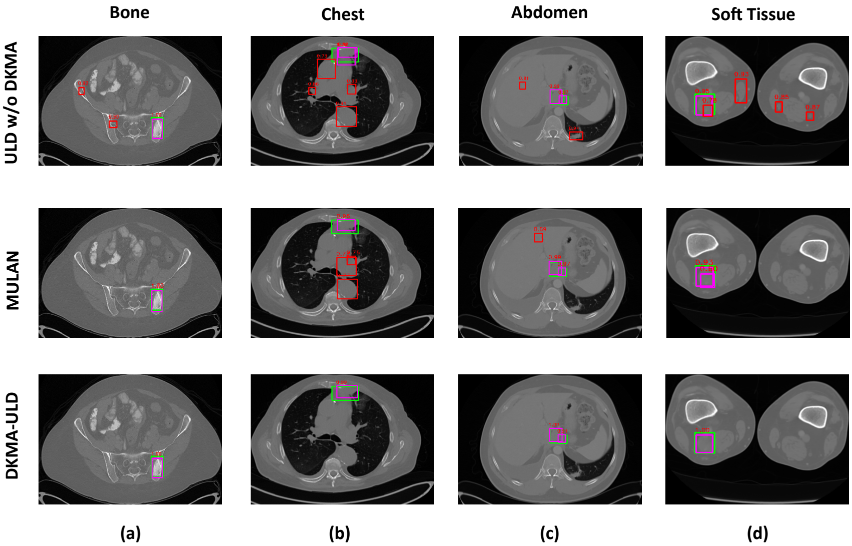


Figure 5: Qualitative comparison of *DKMA-ULD* and MULAN [34] (at FP =2) on CT-scans of different body regions. The green, magenta, and red color boxes represent ground-truth, true-positive (TP), and false-positive (FP) lesion detection, respectively. Please note that ULD w/o DKMA represents when 3 slices with only one HU window ($[1024, 4096]$), default anchors, and without convolution augmented multi-head attention feature fusion are used. We can observe that after incorporating domain knowledge in the form of multi-intensity CT slices, custom anchors, and multi-head attention (i.e., *DKMA-ULD*), the number of FP reduced drastically resulting in improved lesion detection performance as compared to MULAN.

multi-head attention-based feature fusion have enabled *DKMA-ULD* to learn robust representations. Hence, it validates our claim that domain knowledge can alleviate the requirement of a set of diverse datasets for learning good representations in medical imaging analysis. Furthermore, we observe that previous methods such as MULAN [34] and MELD [30] improved the performance of their base models by incorporating techniques such as the use of tagging in MULAN; Missing Annotation Matching (MAM) and Negative Region Mining (NRM) in MELD (refer row 5 and 7 of Table 1). Hence, we also experimented by initializing our network with self-supervised weights. As shown in Table 1, it leads to a gain in performance and achieves a final average sensitivity of 87.16%. For more results on organ-wise sensitivity, refer supplementary material.

Next, we show a comparison of sensitivity at $FP = 4$ for different lesion sizes and average sensitivity (over $FP = \{0.5, 1, 2, 4\}$) for different organs. We observe from the Figure 3(a) that *DKMA-ULD* improves the detection of very small (< 10mm) and medium-sized ($10-30$mm) lesions over 3DCE [32], improved RetinaNet [37], and MULAN [34]. Unlike 3DCE [32] and improved RetinaNet [37], sensitivity values of MULAN [34] for different lesion sizes and organs are not mentioned in their paper. Moreover, as the best trained model of MULAN with tags is not available publicly, we use the official released base model of MULAN [34] (average sensitivity of 84.33%) for computing lesion size-wise and organ-wise sensitivity values on DeepLesion test-set. For a fair comparison, we use the base model of our proposed DKMA-ULD without self-supervision (average sensitivity of 86.88%) in Figure 3 and Figure 4. Further, in Figure 3(b), we observe that our proposed method of including domain-specific information in the lesion detection network improves the average

Sr. No.	HU windows	Backbone	Attention	Custom Anchors	Avg. Sensitivity
1	1	x101			77.59
2	3	x101			80.66
3	5	x101			82.37
4	5	x101	✓		83.30
5	5	x101	✓	✓	84.23
6	5	x152	✓	✓	84.85
7*	**5**	**x152**	✓	✓	**86.88**

Table 2: Ablation studies and average sensitivity comparison (%) of introducing different modules in the proposed lesion detection (*DKMA-ULD*) on the test-set of the DeepLesion dataset. *CT-slices after cropping of outer black-region are used for this experiment.

sensitivity across all organs. Furthermore, we provide a comparison of average sensitivity of MULAN and DKMA-ULD for different lesion-sizes and organs in Figure 4 and demonstrate that DKMA-ULD substantially improves over MULAN in all the cases. For all the above experiments, we use cropped CT slices by clipping the black border region, as mentioned in previous state-of-the-art methods [30, 34], to focus on the region of interest.

Now, we present the ablation study on the introduction of different modules in the proposed lesion detection pipeline, as shown in Table 2. Our proposed 5 HU windows to give organ-specific domain knowledge results in an improvement of approximately 2% in the average sensitivity (82.37%), as shown in row 3 of Table 2. Subsequent to this, we experiment with the inclusion of our novel convolution augmented multi-head attention module for feature fusion and custom anchors to detect varied-sized lesions effectively. We observe a performance boost by achieving an average sensitivity of 84.23%. All the ablation experiments are performed on CT-slices without applying cropping during the pre-processing step. Later in our experiments, we replace our feature extraction backbone with ResNeXt-152 and clip black borders in CT slices enabling the network to focus only on the region of interest. This resulted in a quantitative improvement by achieving a state-of-the-art average sensitivity of 86.88%. Finally, we show a qualitative comparison of lesion detection performance of our proposed *DKMA-ULD* in the form of reduction of FP in Figure 5. For more detailed experimental results, please refer supplementary material.

5 Conclusion and Future Work

In this paper, we demonstrate the potential of exploiting domain knowledge in medical imaging data for developing a robust universal lesion detection network named *DKMA-ULD* that detects lesions across multiple organs of interest with improved sensitivity as compared to the state-of-the-art methods. We also prove that domain-specific weight initialization of ULD using self-supervision on the DeepLesion dataset gives a boost in lesion detection performance. Further, we provide evidence for our choices in using multiple HU windows, lesion-specific custom anchors, multi-head attention-based feature fusion and surpass the state-of-the-art ULD methods by achieving a sensitivity of 87.16% with only 3 slices per patient. This idea of exploiting maximum information in the dataset and learning self-supervised weights can prove useful for any medical image analysis task. In the future, we intend to extend this work by developing an anchor-less lesion detection network and making it robust to shifts in the domain, acquisition protocols, etc. using domain adaptation techniques.

References

[1] CT Lymph Nodes dataset, “The cancer imaging archive (TCIA) public access". https://wiki.cancerimagingarchive.net/display/Public/CT+Lymph+Nodes, 2016.

[2] Kyongtae T Bae et al. CT depiction of pulmonary emboli: display window settings. *Radiology*, 2005.

[3] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3286–3295, 2019.

[4] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *CoRR*, abs/2002.05709, 2020. URL https://arxiv.org/abs/2002.05709.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, D Sargent, Robert Ford, Janet Dancey, S Arbuck, Steve Gwyther, Margaret Mooney, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009.

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Sang-gil Lee, Jae Seok Bae, Hyunjae Kim, Jung Hoon Kim, and Sungroh Yoon. Liver lesion detection from weakly-labeled multi-phase CT volumes with a grouped single shot multibox detector. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 693–701. Springer, 2018.

[13] Zihao Li et al. MVP-Net: Multi-view FPN with position-aware attention for deep universal lesion detection. In *MICCAI*, pages 13–21. Springer, 2019.

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[15] Geert Litjens et al. A survey on deep learning in medical image analysis. *Med. Image Anal.*, 2017.

[16] Kushagra Mahajan, Monika Sharma, and Lovekesh Vig. Meta-DermDiagnosis: Few-Shot Skin Disease Identification using Meta-Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3142–3151, 2020. doi: 10.1109/CVPRW50498.2020.00373.

[17] Samira Masoudi et al. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. *J. Med. Imaging*, 8(1):010901, 2021.

[18] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1): 1–7, 2018.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[20] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Medical image analysis*, 42:1–13, 2017.

[21] Satya P Singh, Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan, and Balázs Gulyás. 3D deep learning on medical images: a review. *Sensors*, 20(18): 5097, 2020.

[22] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4): 341–359, 1997.

[23] You-Bao Tang et al. ULDor: a universal lesion detector for CT scans with pseudo masks and hard negative example mining. In *ISBI 2019*, pages 833–836. IEEE, 2019.

[24] Qingyi Tao, Zongyuan Ge, Jianfei Cai, Jianxiong Yin, and Simon See. Improving deep lesion detection using 3D contextual and spatial attention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 185–193. Springer, 2019.

[25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[26] Äaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[28] Bin Wang, Guojun Qi, Sheng Tang, Liheng Zhang, Lixi Deng, and Yongdong Zhang. Automated pulmonary nodule detection: High sensitivity with few candidates. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 759–767. Springer, 2018.

[29] Xudong Wang, Shizhong Han, Yunqiang Chen, Dashan Gao, and Nuno Vasconcelos. Volumetric attention for 3D medical image segmentation and detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 175–184. Springer, 2019.

[30] Ke Yan, Jinzheng Cai, Adam P Harrison, Dakai Jin, Jing Xiao, and Le Lu. Universal Lesion Detection by Learning from Multiple Heterogeneously Labeled Datasets. *arXiv preprint arXiv:2005.13753*, 2020.

[31] Ke Yan, Jinzheng Cai, Adam P Harrison, Dakai Jin, Jing Xiao, and Le Lu. Universal Lesion Detection by Learning from Multiple Heterogeneously Labeled Datasets. *arXiv preprint arXiv:2005.13753*, 2020.

[32] Ke Yan et al. 3D context enhanced region-based convolutional neural network for end-to-end lesion detection. In *MICCAI*, pages 511–519. Springer, 2018.

[33] Ke Yan et al. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *J. Med. Imaging*, 5(3):036501, 2018.

[34] Ke Yan et al. MULAN: multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In *MICCAI*, pages 194–202. Springer, 2019.

[35] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. 2020. https://d2l.ai.

[36] Wenshuai Zhao, Dihong Jiang, Jorge Peña Queralta, and Tomi Westerlund. MSS U-Net: 3D segmentation of kidneys and tumors from CT images with a multi-scale supervised U-Net. *Informatics in Medicine Unlocked*, 19:100357, 2020. ISSN 2352-9148. doi: https://doi.org/10.1016/j.imu.2020.100357. URL https://www.sciencedirect.com/science/article/pii/S2352914820301969.

[37] Martin Zlocha, Qi Dou, and Ben Glocker. Improving RetinaNet for CT lesion detection with dense masks from weak RECIST labels. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 402–410. Springer, 2019.

[38] Martin Zlocha et al. Universal Lesion Detector: Deep Learning for Analysing Medical Scans. 2019.