



From single to universal: tiny lesion detection in medical imaging

Yi Zhang¹ · Yiji Mao¹ · Xuanyu Lu¹ · Xingyu Zou¹ · Hao Huang² · Xinyang Li¹ · Jiayue Li³ · Haixian Zhang¹

Accepted: 7 April 2024 / Published online: 4 July 2024
© The Author(s) 2024

Abstract

Accurate and automatic detection of tiny lesions in medical imaging plays a critical role in comprehensive cancer diagnosis, staging, treatment, follow-up, and prognosis. Numerous deep learning (DL) based studies have focused on single lesions, providing highly sensitive identification and precise localization. On the other hand, some recent studies have started to concentrate on whole-body lesions, as they could provide systemic clinical support. This paper presents a single-to-universal review of DL studies on tiny lesion detection in medical imaging, with a particular emphasis on detection models and techniques, as well as the data-related aspects such as modality, dimension, and dataset. A wide range of tasks are covered, including traditional single lesion detection tasks such as lung nodules, breast masses, thyroid nodules, and diseased lymph nodes, as well as the emerging task of universal lesion detection. Moreover, the paper provides in-depth analysis and discussion, with the hope of inspiring future research endeavors.

Keywords Artificial intelligence · Deep learning · Lesion detection · Medical imaging

1 Introduction

According to the international agency for research on cancer (IARC) of the world health organization (WHO), there were approximately 19.29 million new cancer cases and 9.96 million cancer-related deaths globally in 2020 (Sung et al. 2021). The incidence and mortality rates of cancer have been steadily increasing each year, significantly impacting the life quality and overall health of individuals worldwide (Sung et al. 2021).

Tiny lesions are of significant importance in cancer diagnosis, staging, treatment, follow-up, and prognosis. Firstly, they serve as early indicators for cancer diagnosis, as detecting tiny lesions can signify the presence of early-stage cancer, facilitating timely diagnosis. Secondly, they contribute to cancer staging by helping determine the extent and severity of the disease based on tumor size, invasion, and metastasis. Thirdly, they are pivotal in guiding cancer treatment, as precise localization and assessment aids in planning surgical resection or radiotherapy. Moreover, tiny lesions are essential for

Yiji Mao, Xuanyu Lu, Xingyu Zou have contributed equally to this work.

Extended author information available on the last page of the article

follow-up and prognosis evaluation. They can serve as markers or biomarkers to assess treatment response and predict patient outcomes. Monitoring changes in lesions during follow-up enables earlier identification of recurrence or metastasis. Overall, tiny lesions are not merely pathological manifestations but also crucial indicators throughout the various stages of cancer.

Medical imaging technology allows for non-invasive and convenient acquisition of high-resolution images. However, the accurate detection of tiny lesions within complex structural backgrounds in these images remains a labor-intensive task that heavily relies on the expertise of radiologists. Even experienced radiologists may still overlook some of these tiny lesions, potentially leading to delayed diagnosis and depriving patients of their best chance for appropriate treatment.

Automated tiny lesion detection shows promise in improving this situation and is indispensable in clinical workflows nowadays. It involves the use of algorithms and models to accurately and efficiently detect lesions in medical images, enabling early identification of potential cancer risks. In recent years, the remarkable achievements of deep learning (DL) have contributed to the growing interest in this field within medical imaging. Traditional knowledge-based algorithms have gradually been replaced by DL models due to their powerful ability for recognizing and localizing complex patterns and extracting higher-level feature representations.

Lesions closely related to high-risk cancers, such as lung nodules, breast masses, etc., receive primary attention. Numerous single lesion detection (SLD) models have been proposed. These models focus on specific categories of lesions and are capable of providing highly sensitive detection for target lesions.

However, SLD alone is insufficient to provide comprehensive clinical support. In clinical practice, it is commonly observed that patients have multiple lesions, and these lesions are often interrelated (Yan et al. 2018a). Detecting all of them can support doctors in integrating diverse lesion characteristics and information, enabling them to generate more comprehensive and accurate diagnoses, as well as formulate personalized treatment plans. Furthermore, compared to using an unknown number of SLD models to detect any potentially existing lesions, detecting all lesions in one-shot is clearly more efficient and better aligned with clinical workflow. Therefore, the universal lesion detection (ULD) holds significant clinical value as it enables the detection of multiple lesions throughout the body using a single model, making it a prominent and rapidly evolving field.

In this paper, we review lesion detection from single to universal using medical imaging. Specifically, we overview and discuss the recent advancements in four popular SLD tasks: lung nodules, breast masses, thyroid nodules and diseased lymph nodes detection. Additionally, the ULD task is included. Various imaging modalities, including computed tomography (CT), X-ray, magnetic resonance imaging (MRI), and ultrasound, are considered. For each task, our main focus lies in analyzing the innovations of the representative studies and clarifying their relationship, as well as the performance comparison. We also introduce the datasets, data modalities, and data dimensions used in each study. Furthermore, drawing upon our experience of artificial intelligent (AI) in medical imaging, we provide an in-depth discussion and analysis of lesion detection area, along with possible improvements for current challenges and future directions. We hope this paper could offer readers a better understanding and minor inspiration on both medical and technical aspects of tiny lesion detection.

Please note that the term “detection” mentioned in this paper refers to the task of object detection in the field of computer vision.

1.1 Search strategy and selection criteria

The bibliographic literature was thoroughly searched to identify relevant studies. The search was conducted on various platforms, including Google Scholar, Web of Science, Pubmed, and the Institute of Electrical and Electronics Engineers (IEEE). The following keywords are used in cross combinations to ensure the integrity of literature collection: “detection,” “deep learning,” “CNN,” “nodule,” “mass,” “node,” and “lesion.” Additionally, we conducted a combined search using the names of dozens of lesions along with “detection”. The search was limited to English papers published between 2015 and 2023. To ensure the quality of the selected papers, we manually chose groundbreaking, highly-cited, and high-level papers. Initially, we collected 2260 papers and scanned their abstracts, ultimately narrowing it down to 412. Finally, we leave 98 papers after carefully reading the content and code (if provided).

1.2 Prototype models

Among the detectors we investigated, we have observed the presence of many classic models. We categorized seven models that have been used two or more times and referred to them as prototype models. A sketch and a brief introduction are provided for each of them.

1.2.1 U-Net

U-Net (Ronneberger et al. 2015), which was proposed by Ronneberger et al. in 2015, was originally developed to tackle the semantic segmentation challenges encountered in medical image analysis. It is named after its U-shaped architecture. However, the definition of U-Net has become quite broad nowadays. Many models with a U-shaped design are commonly referred to as “U-Net”, “U-Net-based” or “U-Net-like”. This convention is followed in this paper as well.

U-Net adopts an encoder-decoder structure. The encoder (downsampling path) gradually reduces the size of and enlarges the number of channels of feature maps through a series of convolutional layers and pooling layers, extracting high-level semantic information while reducing computational complexity. The decoder (upsampling path) restores the size and number of channels of feature maps through a series of upsampling layers and feature fusion operations, symmetrically connecting with the encoder to reconstruct detail and positional information. Skip connections connect the feature maps from the encoder to the corresponding levels in the decoder, allowing the preservation and fusion of features at different resolutions. This enables the network to retain detail information while handling coarse semantic features. It is obvious that this encoder-decoder structure is highly compatible with the design pattern of detection models: a backbone for feature extraction, followed by a neck for feature aggregation. Therefore, U-Net is not only popular in segmentation tasks but also widely used in many detection tasks, especially in medical imaging tasks. A simple U-Net-based detector is depicted in Fig. 1.

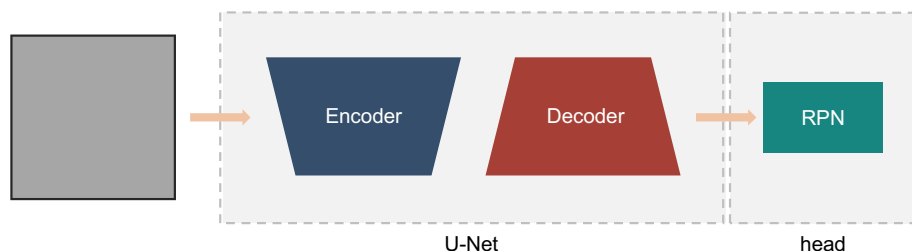


Fig. 1 The sketch of a U-Net-based detector

1.2.2 You only look once (YOLO)

YOLO (Redmon et al. 2016) was proposed by Joseph Redmon et al. in 2016. The core idea of YOLO is to transform the object detection problem into a regression problem for a single neural network, which predicts both the object class and bounding box in a single forward pass.

YOLO is a one-stage detector designed for real-time object detection with a balance between accuracy and speed. Its key features include: end-to-end regression, enabling multi-scale detection by dividing the input image into a grid, and direct output of bounding box coordinates and class probabilities for detected objects. Figure 2 illustrates the structure of YOLOv5, which detects objects across three scales of feature maps.

YOLO offers the benefits of high speed and real-time performance, making it well-suited for applications that require fast object detection. Today, YOLO has become a series, with iterative developments resulting in the latest versions, including YOLOv8. In the context of lesion detection tasks, YOLOv3 (Redmon and Farhadi 2018) and YOLOv5 are among the most commonly utilized versions.

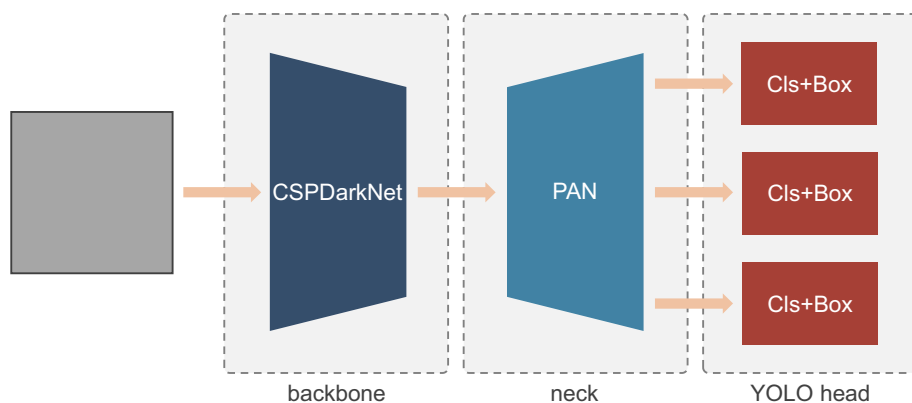


Fig. 2 The sketch of YOLOv5 network

1.2.3 Single shot multibox detector (SSD)

SSD was proposed by Liu et al. (2016), at around the same time as YOLOv1, and is also a one-stage detector designed for achieving efficient object detection. However, unlike YOLOv1, which predicts object classes and bounding boxes directly on a single global feature map, SSD utilizes distinct detection heads to detect objects of different sizes on 6 scales of feature maps. This approach has become a paradigm followed by many subsequent detectors, including YOLOv3/v5. SSD is illustrated in Fig. 3.

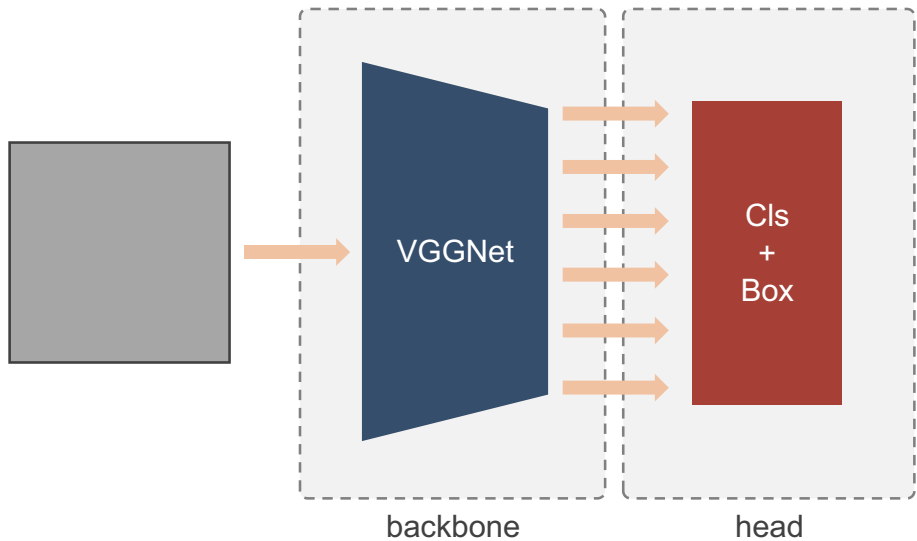


Fig. 3 The sketch of SSD network

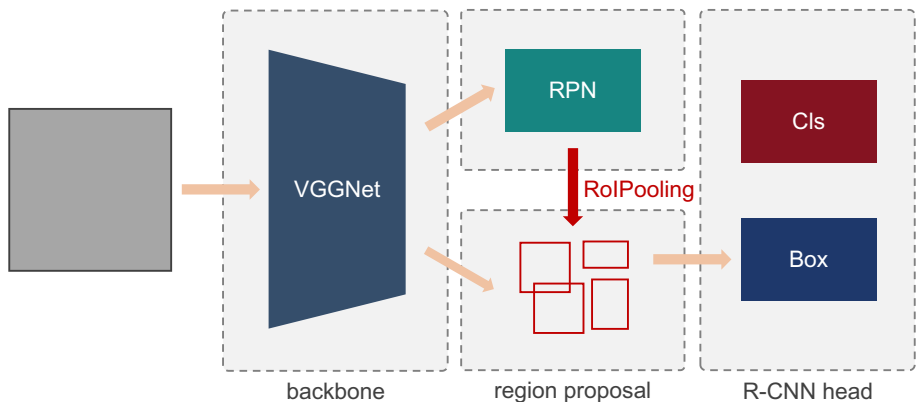


Fig. 4 The sketch of faster R-CNN network

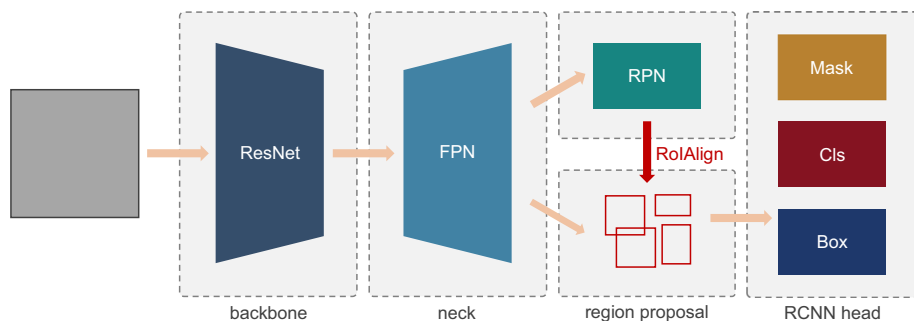


Fig. 5 The sketch of Mask R-CNN network

1.2.4 Faster R-CNN

Faster R-CNN (Fig. 4) is a two-stage object detection model proposed by Ren et al. (2015). It is an improved version of the R-CNN series models, aiming to enhance the accuracy and speed of Fast R-CNN (Girshick 2015).

Faster R-CNN employs end-to-end training and jointly optimizes the RPN head and the R-CNN head to enhance the performance through mutual interaction. This approach has achieved significant advancements in the field of object detection, offering a valuable combination of accuracy and speed. By incorporating RPN as a candidate box generator, Faster R-CNN enables more precise and efficient object detection. Its two-stage structure makes Faster R-CNN well-suited for the logical process of initially detecting lesions and subsequently eliminating false positives. As a result, Faster R-CNN has been widely adopted in the field of medical lesion detection.

1.2.5 Mask R-CNN

Mask R-CNN is an advanced model for object detection and instance segmentation, proposed by He et al. (2017). As shown in Fig. 5, it is an extension of Faster R-CNN that not only identifies and locates objects but also generates segmentation masks.

In terms of object detection, the primary distinction between Mask R-CNN and Faster R-CNN lies in the approach used to extract RoI features. The proposed RoIAlign layer, which performs a similar function as the RoIPooling in Faster R-CNN, utilizes bilinear interpolation to circumvent the need for two quantization operations in RoIPooling. This approach effectively mitigates information loss.

Mask R-CNN has made significant advancements in object detection and instance segmentation, providing crucial information for further image understanding and analysis. This paper categorizes the models following the two-stage structure and utilizing RoIAlign to extract RoI features as ‘Mask R-CNN’ series.

1.2.6 RetinaNet

RetinaNet (Lin et al. 2017b) is a one-stage multi-scale detector introduced by Facebook AI Research in 2017. It addresses two fundamental challenges in object detection: class

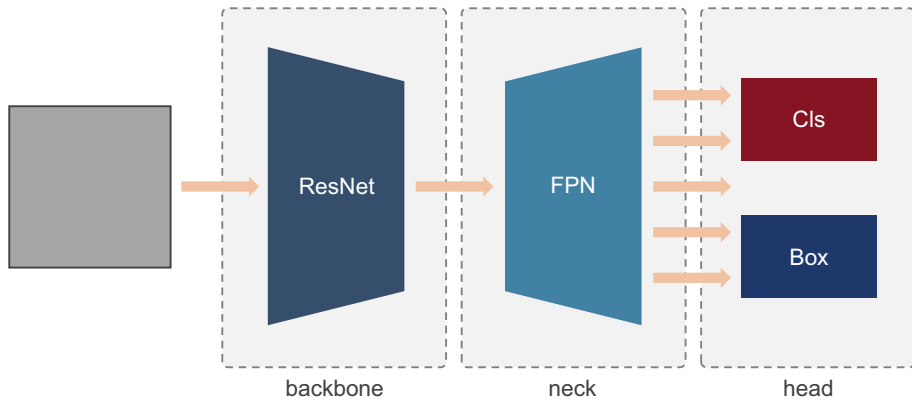


Fig. 6 The sketch of RetinaNet

imbalance and small object detection. It has achieved remarkable advancements in terms of both performance and efficiency.

The core idea of RetinaNet is the introduction of a loss function called focal loss, aimed at addressing the problem of class imbalance in object detection. The conventional cross-entropy loss often faces challenges due to the presence of a substantial number of background class samples, making it difficult to effectively learn features of rare object classes. Focal loss adjusts the sample weights, prioritizing the training of difficult-to-classify samples, thereby enhancing the performance of object detection.

The architecture of RetinaNet is shown at Fig. 6. It adopts the FPN as the neck, leveraging feature maps at different levels to provide multi-scale information. RetinaNet incorporates classifiers and regressors on each feature layer to predict bounding boxes and categories at various scales. This enables effective and accurate object detection, especially small-sized objects. RetinaNet has gained wide popularity and is considered a powerful model in object detection tasks.

1.2.7 CenterNet

In 2019, Wang et al. proposed CenterNet (Duan et al. 2019). Similar to CornetNet (Law and Deng 2018) and ExtremeNet (Zhou et al. 2019), CenterNet is also a keypoints-based anchor-free detector, pioneering the field of anchor-free detectors. Its core idea is to detect objects by predicting their center points, which leads to robust and fewer FP detection, as well as faster inferring speed.

As shown in Fig. 7, the network architecture of CenterNet consists of a backbone network and several task-specific sub-networks. The backbone, typically based on ResNet (He et al. 2016) or Hourglass network, is responsible for extracting features from the input images. The sub-networks are responsible for predicting the heatmap of center point, the bounding box sizes, and the center point offsets.

CenterNet transforms the object detection problem into a regression task by predicting the positions of object center points on the feature map. During inferring, CenterNet utilizes Gaussian heatmaps to represent the center points of objects. Such anchor-free models are applied in lung nodule detection and ULD.

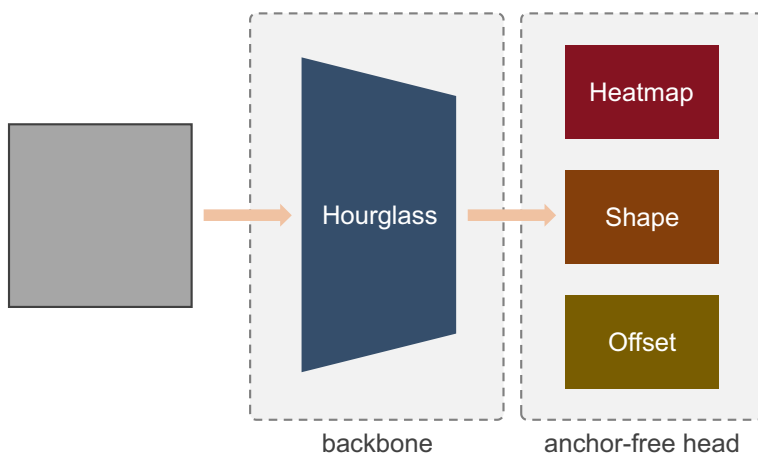


Fig. 7 The sketch of CenterNet

1.3 Input data

Natural image is typically in a three-channel 2D format, with a shape like $(3, H, W)$, where ‘3’ denotes the RGB channels, ‘H’ refers to the height of the image, and ‘W’ represents the width of it. Unlike natural image, medical image often consist of single-channel 2D slices or 3D scans due to their unique imaging modalities. Additionally, in order to reduce memory and computation cost, some studies have adopted the practice of decomposing 3D image data into 2D, 2.5D or $(2+1)D$ forms, allowing for the retention of volumetric information to a certain extent.

1.3.1 2-dimension (2D)

Medical imaging obtained using techniques such as X-ray, MRI, etc., are typically 2D. These images have only one channel that display the density, brightness, and texture characteristics of different tissues and abnormalities. As shown in Fig. 8a, the shape of such images are $(1, H, W)$, where ‘1’ denotes the gray scale channel, making they can be trained using 2D models just like natural images.

1.3.2 3-Dimension (3D)

CT scans acquire a substantial amount of X-ray projection from various angles and utilize a reconstruction algorithm to restore the three-dimensional structure of the target. As a result, CT scans are often volumetric data. In addition, modern ultrasound technology has also advanced to include 3D imaging. 3D images can provide comprehensive anatomical and pathological information. Therefore, as shown in Fig. 8b, 3D models directly employ the volumetric data as input, with the shape of $(1, D, H, W)$, where ‘D’ represents the number of axial slices.

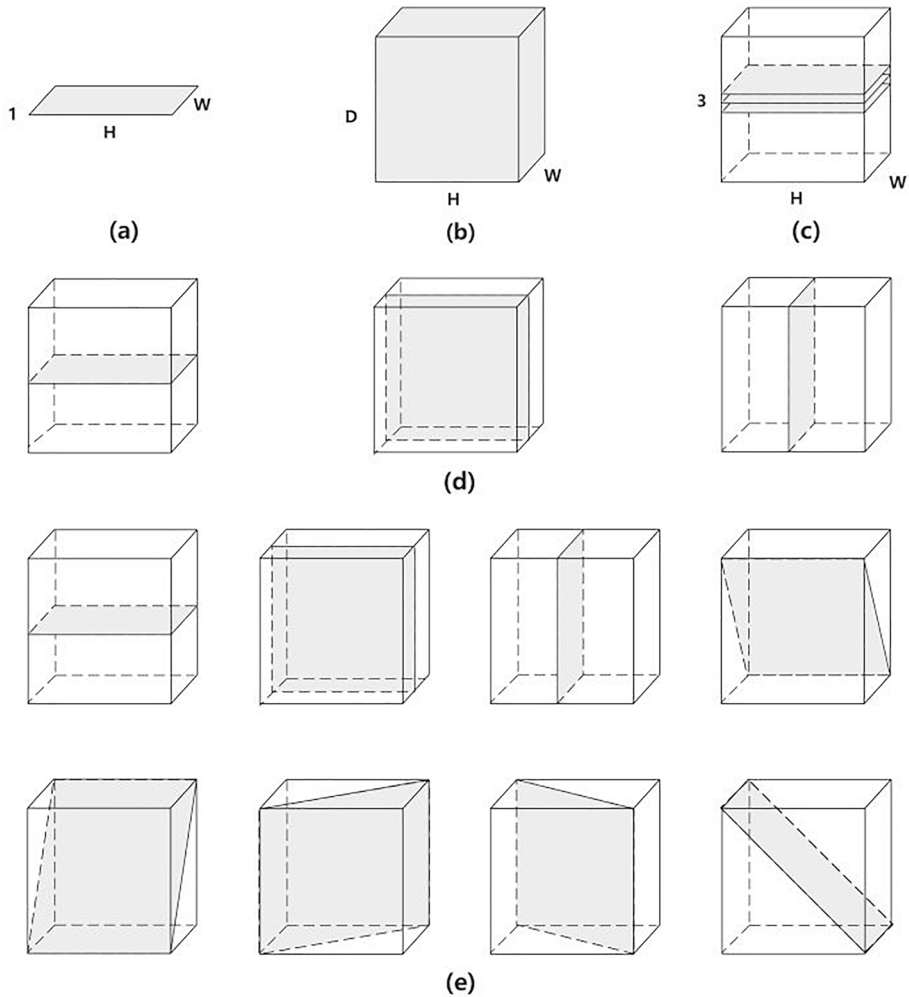


Fig. 8 Input data of different dimensions

1.3.3 (2+1)-Dimension ((2+1)D)

For the majority of single lesion datasets, such as LUNA16 (Setio et al. 2017), there are often only hundreds of available CT scans. In today's context, a model that relies fully on 3D convolution is computationally feasible. However, when dealing with much larger datasets like DeepLesion (Yan et al. 2018a), or multiple datasets like Yan et al. (2020), tens of thousands of CT scans are included, rendering pure 3D convolution impractical. Therefore, the (2+1)D input, with the shape of (1, 3, H, W), is introduced (shown in Fig. 8c). Every three consecutive CT slices are combined to create a three-channel image, enabling the training and inferring of CT scans using 2D models that have been pretrained on ImageNet (Krizhevsky et al. 2012).

1.3.4 2.5-Dimension (2.5D)

CT scans contain ample contextual information, but the utilization of 3D data for model training poses challenges due to computational limitations. Consequently, some studies have sliced the 3D CT scans from various views, generating multiple 2D slices. Figure 8 illustrates the views of (d) axial, sagittal, and coronal slices; and (e) additional diagonal slices. These planes are spliced together in channel dimension to form a multi-channel image, with the shape of (1, N, H, W), where 'N' denotes the number of slices. This approach partially preserves spatial information while considerably reducing the computational burden, and was prevalent in studies on false positive reduction (FPR_e). However, with advancements in hardware devices in recent years, this approach is no longer as common, as it is now possible to work with full 3D data.

1.4 Evaluation metrics

In this subsection, the evaluation metrics used in this paper are introduced.

1.4.1 Accuracy

Accuracy (ACC) is a commonly used evaluation metric for classification and detection tasks, measuring the proportion of correctly predicted object categories out of the total predicted objects. ACC can be calculated by the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where *TP* represents true positives, *TN* represents true negatives, *FP* represents false positives and *FN* represents false negatives. However, ACC has some limitations:

- Not suitable for imbalanced datasets: ACC may not be accurate when the number of samples in one class significantly outweighs the others. For example, in lesion detection, if the number of negative samples (tissue background) is much larger than positive samples (target lesions), ACC may overly reflect model performance.
- Unable to capture error types: ACC cannot differentiate between FPs and FNs. In lesion detection, missing a detection (FN) may be more critical than false alarms (FP), but ACC fails to provide such information.
- Limited for multi-class problems: ACC is not suitable for evaluating multi-class object detection tasks as it cannot distinguish specific class errors.

Therefore, to comprehensively and accurately evaluate the performance, ACC is often used in conjunction with other evaluation metrics, such as recall, precision, specificity, etc. These metrics provide more specific evaluation and help us understand model performance in different aspects.

1.4.2 Recall/sensitivity/TPR

Recall (REC), also known as sensitivity or true positive rate (TPR), quantifies the model's ability to correctly identify positive instances among all actual positive instances in the dataset. REC can be defined as:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

A high REC indicates the model's proficiency in correctly capturing a substantial portion of positive instances. Note that REC is intricately related to the FPs at a specific threshold. When striving for higher REC, it becomes more probable that the number of FPs may increase as well. Therefore, it is meaningful to consider REC with the FP-related metrics.

1.4.3 Specificity/TNR

Specificity (SPEC) or true negative rate (TNR) measures the accuracy of a binary classification model in identifying negative instances. It evaluates the proportion of accurately predicted negative samples in the actual negative samples. SPEC can be defined as:

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

The SPEC value ranges from 0 to 1. Higher SPEC values indicate a stronger ability of the model to discriminate negative instances. In certain cases, the ability to correctly identify negatives may be more important than the ability to identify positives. For example, in medical screening for early disease detection, a high SPEC model is needed to minimize the risk of FNs.

1.4.4 Precision

Precision (PREC), also referred as positive predictive value, measures the accuracy of the model's positive predictions. It assesses the proportion of true positive predictions among all the positive predictions made by the model. PREC is calculated as:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

A high PREC value indicates that the model excels in making accurate positive predictions. However, it does not account for the missed positive instances. Hence, PREC and REC are commonly used together to assess the performance of a model.

1.4.5 ROC and AUC

The receiver operating characteristic curve (ROC) curve is a graphical representation of TPR and false positive rate (FPR), which illustrates the performance of a binary classification model. The ROC curve can be presented as:

$$ROC(T) = (FPR(T), TPR(T)) \quad (5)$$

where T represents the threshold. The ROC curve is obtained by plotting the TPR against the FPR for various classification thresholds. Each point on the curve represents the trade-off between FPs and TPs at a specific threshold. By examining the shape of the ROC curve and calculating the area under curve (AUC), we can evaluate the model's performance. AUC can be calculated as:

$$\text{AUC} = \int \text{TPR}(\text{FPR}), d(\text{FPR}) \quad (6)$$

The value of AUC ranges between 0 and 1, where a value closer to 1 indicates better performance of the model, and a value closer to 0.5 indicates performance close to random guessing.

1.4.6 Average precision (AP)

AP captures the trade-off between PREC and REC across different classification thresholds. It quantifies the model's performance comprehensively by computing the area under the PREC-REC curve. The AP score is obtained by calculating the average PREC values at different REC levels. Higher AP scores indicate better model performance in maintaining high PREC while achieving high REC. Mathematically, the equation for AP can be expressed as:

$$\text{AP} = \int_0^1 \text{precision}(r), d(r), \quad (7)$$

where $\text{precision}(r)$ is the PREC at a specific REC level r .

Since REC is defined at a specific intersection over union (IoU) value, AP50 or AP0.5 is commonly used to represent the AP value with an IoU threshold of 0.5. Similarly, AP75, AP90, and other variations can be used accordingly. On the other hand, the mean average precision (mAP) represents the mean value of APs of all classes across multiple IoU thresholds, typically ranging from 0.5 to 0.95 with a step of 0.05. In addition, some studies use metrics such as mAP0.5, representing the mAP when the IoU threshold is fixed at 0.5.

1.4.7 FROC/CPM

Free-response receiver operating characteristic (FROC) is a variant of ROC curve that describes the correlation between sensitivity and the number of FPs per scan. It is used to address the issue of ROC not being able to handle multiple abnormalities on a single image. This is because medical image detection tasks require a high REC and can tolerate the existing of multiple FPs.

FROC discusses how high REC can be obtained under the condition of how many FPs. On the other hand, competitive performance measure (CPM) is the average sensitivity at $1/8, 1/4, 1/2, 1, 2, 4, 8, \dots$ FPs/scan, which is the selected breakpoints ($2^k, k \in \{-3, -2, -1, 0, 1, 2, 3, \dots\}$) along the FROC curve. Nowadays, many studies refer to CPM as FROC score. The calculation can be expressed as:

$$\text{CPM} = \frac{1}{n} \sum_{i=1}^n S_i \quad (8)$$

where S_i is the sensitivity at i th breakpoint.

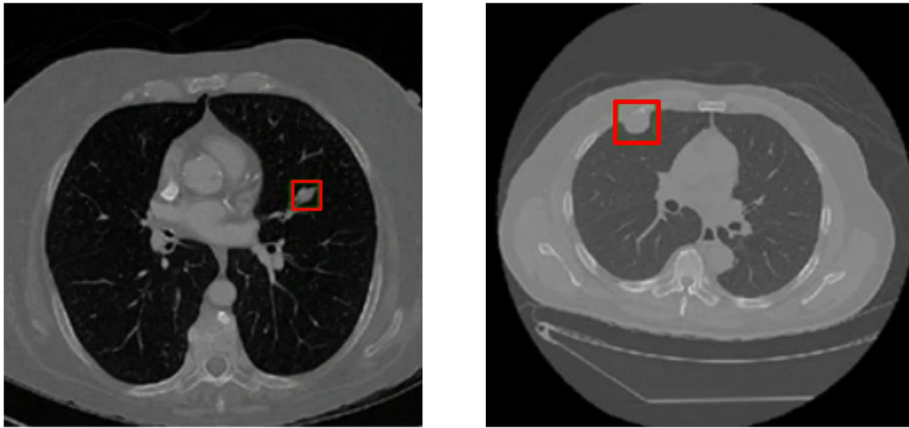


Fig. 9 Lung nodules in LUNA16 dataset

2 Single lesion detection

In the past decade, extensive studies have been conducted on detecting specific types of lesions. The attention given to a lesion detection task is strongly associated with the incidence or malignancy. In this section, we analyze the current progress in detecting four prevalent types of lesions: lung nodules, breast masses, thyroid nodules and diseased lymph nodes. Each type of lesion is discussed in a dedicated subsection.

2.1 Lung nodules

Lung cancer is the leading cause of cancer-related deaths worldwide (Sung et al. 2021). As shown in Fig. 9, lung nodules are small abnormal findings in the lung that often serve as early indicators of lung cancer. The significance of detecting lung nodules lies in regular imaging examinations that enable timely identification and monitoring of changes in these nodules, subsequently determining the need for further diagnosis and treatment. This early screening method assists healthcare professionals in recognizing patients at risk of developing lung cancer, enabling appropriate interventions and improving treatment success and overall survival (Armato III et al. 2011). Therefore, lung nodule detection is crucial for the early diagnosis and treatment of lung cancer.

2.1.1 Datasets

In clinical practice, CT is commonly employed for the identification of lung nodules. It delivers high resolution images, offering multi-planar reconstruction and visualization of soft tissues and bones. Moreover, CT scans exhibit sensitivity to various densities, which facilitates accurate diagnosis and evaluation of lung nodules. Notably, the advancement of lung nodule detection on CT images has been significantly accelerated by competitions like LUNA16 and TianChi, as well as by the public datasets listed in Table 1.

Table 1 Publicly available lung nodule datasets

Dataset	Modality	Scans	Nodules	Availability
LIDC-IDRI	CT	1108	2562	Open access
LUNA16	CT	888	1186	Open access
DSB17	CT	2101	N/A	For participants
PN9	CT	8798	40,439	Need request
TianChi	CT	3000	N/A	1000 for open access 2000 for participants

2.1.1.1 LIDC-IDRI The LIDC-IDRI (lung image database consortium and image database resource initiative) (Armato III et al. 2011) dataset is a publicly available medical imaging dataset widely used for lung nodule detection and characterization research. It consists of 1018 CT scans from 1010 patients, with annotations by four radiologists for the presence and characteristics of lung nodules. The dataset provides a valuable resource for developing and evaluating algorithms and models for lung nodule classification, detection, and segmentation.

2.1.1.2 LUNA16 LUNA16 (Setio et al. 2017) is the dataset used in LUNG Nodule Analysis 2016 challenge, with a collection of 888 CT scans selected from LIDC-IDRI. The cases with slice thickness ≥ 2.5 mm are excluded. Additionally, only nodules with a diameter ≥ 3 mm and annotated by three or more experts are retained.

2.1.1.3 DSB17 The DSB17¹ dataset originated from the Data Science Bowl 2017 competition hosted by Kaggle.² This dataset draws from the resources of the National Cancer Institute and includes 2101 CT scans of high-risk patients. It is worth noting that DSB17 only includes per-subject binary labels indicating whether each case was diagnosed with lung cancer within a year after the scanning. The champion team (Liao et al. 2019) annotated 832 lung nodules boxes by themselves to fit their model.

2.1.1.4 PN9 In comparison with widely used datasets such as LUNA16, PN9 (Mei et al. 2021) presents a significant increase in scale, providing 8798 CT scans and 40,439 annotated nodules that cover the nine most common types of lung nodules, including solid nodules and ground-glass nodules, among others. Moreover, PN9 exhibits a slice thickness ranging from 0.4 mm to 2.5 mm and includes numerous smaller nodules, making it clinically suitable for lung nodule detection.

2.1.1.5 TianChi The Tianchi³ Medical AI Competition 2017 provided 3000 cases of low-dose lung CT images of high-risk patients, along with corresponding labels for the location and diameter of the nodules. All CT images have a slice thickness of < 2 mm, with nodule diameters ranging from 5 to 30 mm. Currently, only the 1000 CT cases used in the preliminary round are publicly available.

¹ <https://www.kaggle.com/c/data-science-bowl-2017>.

² <https://https://www.kaggle.com/>

³ <https://tianchi.aliyun.com/competition/entrance/231724/information>.

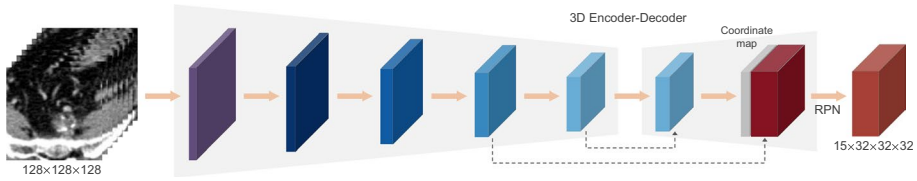


Fig. 10 The detector of LeakyNet

2.1.2 Detectors

2.1.2.1 Beginning from leaky noise-or network Leaky noise-or network (LeakyNet) (Liao et al. 2019), the champion solution of the DSB2017 competition, has laid a solid foundation in terms of code and model, paving the way for numerous subsequent advancements. As shown in Fig. 10, LeakyNet adopts a U-shaped encoder-decoder structure, and utilizes a 3D ResNet as the backbone. The network takes input crops with a size of 128^3 and progressively downsamples the size to $1/16$ of the original before upsampling it to $1/4$ for lung nodule detection. In particular, a normalized coordinate map is appended to the feature map in order to augment the positional information of the crop.

LeakyNet utilizes binary cross-entropy loss to supervise the lung nodule classification, and smooth L1 loss to supervise the bounding box regression. Additionally, considering that lung nodules typically have a circular-like shape, LeakyNet treats the bounding box and anchor as cubes and defines the regression target as a quadruple (d_x, d_y, d_z, d_r) , representing the relative offset to the center point (z, y, x) and the anchor size (r) . The calculation formula is illustrated in Equation (9). LeakyNet uses the same detection head as YOLO (Redmon et al. 2016), combining the class branch and box branch. With three pre-defined anchors, the region proposal network (RPN) outputs a tensor with the shape of $(15, 32, 32, 32)$. Furthermore, the hard negative mining strategy is applied to emphasize the learning of challenging samples. These designs, including the hyperparameters configuration, have become widely adopted defaults.

$$\begin{aligned} d_x &= (G_x - A_x)/A_r, \\ d_y &= (G_y - A_y)/A_r, \\ d_z &= (G_z - A_z)/A_r, \\ d_r &= \log(G_r/A_r). \end{aligned} \quad (9)$$

Based on LeakyNet, DeepSEED (Li and Fan 2020) incorporates the squeeze and excitation (SE) module (Hu et al. 2018), an attention mechanism designed to weigh channel-wise features, into the residual blocks. Furthermore, the focal loss (Lin et al. 2017b) is utilized to help alleviate the imbalance of positive and negative samples. DeepLN (Xu et al. 2020) employs LeakyNet based on the segmentation results of the lung parenchyma. Considering variations in spacing within their private dataset, the authors divided it into two parts: ThinSet and ThickSet. They trained two separate models accordingly and obtained the final predictions by combining the outputs of both models using a weighted non-maximum suppression (NMS). In Han et al. (2022), Han et al. removed the coordinate map and redefined the anchors as $[6, 15, 30]$ to accommodate the diameters of lung nodules. Regarding other related works, Guo et al. and Lin et al. focused on exploring module design and feature

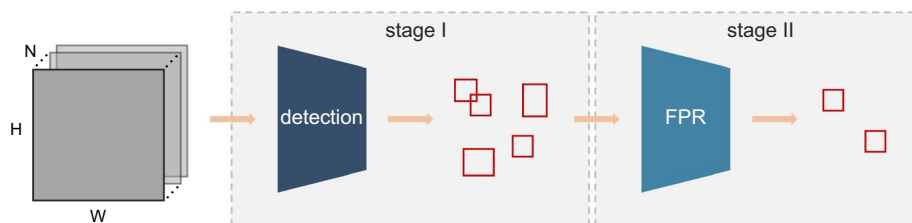


Fig. 11 The pipeline consists of two stages: candidate generation, followed by FPR

interaction, proposing MSANet (Guo et al. 2021) and IR-Unet++ (Lin et al. 2023), respectively. MSANet designed the SCConv module and the ECA module to explore discriminative information, and presented a multi-scale aggregation interaction strategy between the encoder and decoder for integrating multi-level information from adjacent resolutions. Similarly, IR-Unet++ combined SE module and Inception module as the alternative component of each residual block, and employed the U-Net++ (Zhou et al. 2018) structure to communicate feature cross levels.

2.1.2.2 Multi-scale prediction The solution that topped the LUNA16 competition also drew inspiration from LeakyNet. It incorporated FPN (Lin et al. 2017a) to detect candidates on multi-scale feature maps. Additionally, a separate FPRE network has been implemented to eliminate FPs. This two-stage workflow (as shown in Fig. 11), which can be traced back to traditional machine learning era, has become a widely accepted pattern among many DL models.

Zheng et al. (2022) proposed the 3D multi-scale feature extractor (MSFE), constructing the MSFD-Net for generating nodule candidates. The MSFD-Net predicts smaller candidates on feature maps with 1/4 resolution and larger ones on maps with 1/8 resolution. A lightweight MSFE is then used to extract features from these candidates, along with a simple classifier to determine their authenticity as nodules.

3DFPN-HS² (Liu et al. 2019b) generated candidates on four feature maps, (P2, P3, P4, and P5), which guarantees detection across a wide range of scales. Its fully-connected 3D FPN integrates both low-level texture information and high-level semantic information, enhancing the detection of candidate nodules of various sizes. In the FPRE stage, the authors rely on the prior knowledge that true nodules present similar signals and shapes in adjacent slices, while FPs often lack such characteristics. They propose the use of location history image (LHI) to effectively reduce FPs.

Zhao et al. (2023) obtained candidates on three feature maps. During the FPRE stage, three 3D crops, with respective resolution of $24 \times 40 \times 40$, $12 \times 20 \times 20$ and $6 \times 10 \times 10$, were extracted for each candidate. This process considered various nodule sizes, and jointly determined whether the candidates were true nodules.

Introducing additional information may help with the learning of the model. Harsono et al. and Jaeger et al. utilized RetinaNet to develop their detectors, namely I3DR-Net (Harsono et al. 2022) and Retina U-Net (Jaeger et al. 2020) respectively. I3DR-Net incorporates the Inflated 3D ConvNet (Carreira and Zisserman 2017) to construct each convolutional block. Additionally, taking into account the similarities in input data shape between video data and 3D medical data, I3DR-Net is pretrained on the video action recognition dataset kinetics (Carreira and Zisserman 2017). On the other hand, Retina U-Net incorporates supplementary segmentation supervision. The top-down

path integrates additional pyramid levels, P1 and P0, to enable a U-shape segmentation branch. Experiments demonstrate that the inclusion of segmentation supervision improves detection accuracy.

2.1.2.3 R-CNN-based While LeakyNet contributes a lot to nodule detection, Faster R-CNN is another path to achieve this. To the best of our knowledge, Ding et al. (2017) were the first to utilize a Faster R-CNN-based detection model in the LUNA16 challenge, enabling them to maintain the top position for a period of time. Their model follows the two-stage workflow, as shown in Fig. 11, and takes (2+1)D images as the input of a 2D VGG-16 backbone. To accommodate the size of lung nodules, an additional deconvolution layer is added at the end to enlarge the resolution of the detection layer. In the FPRE stage, a $20 \times 36 \times 36$ crop is extracted and processed through a 3D ConvNet for nodule discrimination. This “2D detection plus 3D FPRE” pipeline strikes a balance between performance and computational efficiency, and has gained widespread adoption in their contemporaneous studies. A similar work can be found in Huang et al. (2019).

Building upon the work of Ding et al. (2017), Sun et al. (2018), Xie et al. (2019) and Wang et al. (2018b) made further improvements. Sun et al. (2018) and Xie et al. (2019) introduced an additional RPN network to the shallow feature map to generate one more set of candidates. This can be beneficial for detecting small nodules due to the smaller receptive field and preservation of more texture details in the shallow map. In the FPRE stage, the network takes 2.5D data as input, consisting of 9 slices, each of size 35×35 . In particular, the authors employed a boosting strategy, dividing the candidate set into three subsets of s1, s2, and s3. The misclassified samples in s1 are merged into s2, and the ones in s2 are merged into s3, resulting in hierarchical accuracy improvement for difficult samples. The FP identification is decided by a three-model vote. On the other hand, Wang et al. (2018b) detected candidates on three feature scales. The FPs are distinguished based on the concatenation of the candidate crop and the segmentation mask generated by a 3D U-Net.

3D models have been increasing in popularity as computing power has advanced in recent years. In studies such as DeepLung (Zhu et al. 2018), NoduleNet (Tang et al. 2019c, b), etc, 3D crops with resolution of $96 \times 96 \times 96$ or $128 \times 128 \times 128$ are commonly used to provide detailed information. While DeepLung (Zhu et al. 2018; Tang et al. (2019b) can almost be considered as the original 3D Faster R-CNN, NoduleNet stands out with its integrated design that combines segmentation, detection, and FPRE tasks in the end-to-end structure. Specifically, NoduleNet used a backbone similar to LeakyNet to detect candidate nodules. However, considering that the low-level features contain richer detail information, the RoI features for FPRE are extracted from the shallow feature map from the top-down path instead of the usual practice, that is, the deep feature map from bottom-up path.

NoduleNet (Tang et al. 2019c) enlightens the work of SANet (Mei et al. 2021). SANet introduced the slice grouped non-local (SGNL) module, a variant of non-local module (Wang et al. 2018a), into the top-down path of NoduleNet. The SGNL module is capable of capturing cross-channel information and modeling long-range dependencies among positions and channels. Additionally, SANet further increased the resolution of the detection layer, reaching half the size of the input image. Building upon SANet, Xu et al. propose LSSANet (Xu et al. 2022). The main difference lies in the replacement of the SGNL module with the long short slice grouping (LSSG) module, which alleviates the computational and spatial burden of non-local module.

While the majority of studies have been conducted using CT images, Tsai and Peng (2022) utilized 21,189 2D X-ray images from the National Institute of Health (NIH)

chest X-ray dataset. They proposed a dual head network (DHN), a hybrid model that combines Faster R-CNN and Mask R-CNN. The main distinction of DHN is its two-branch detection head: a global head predicts a binary label indicating the presence of nodules, and a local head regresses bounding boxes of detected nodules. Moreover, a dual head data augmentation strategy is introduced to avoid partial optimization of one head.

2.1.2.4 SSD- and YOLO-based Ma et al. introduced a SSD-based detector called the group-attention SSD (GA-SSD (Ma et al. 2019). This detector employs ResNeXt (Xie et al. 2017) as its backbone to form FPN-like layers for detecting objects at multiple scales. The proposed GA module allows the model to selectively amplify specific channels that require greater attention.

Some studies are built upon YOLO series. George et al. (2018) and Zhou et al. (2022) utilized a 2D YOLOv1 and a 2D YOLOv5 detector, respectively. Huang et al. proposed a variant of 3D YOLOv3 (Redmon and Farhadi 2018), known as OSAF-YOLOv3 (Huang et al. 2022). They introduced the OSAv2 module to filter out redundant features and prevent gradient vanishing, as well as the feature fusion scheme to address the issue of information loss resulting from upsample-based feature fusion.

2.1.2.5 Anchor-free detectors In Liu et al. (2021b), Liu et al. proposed an end-to-end framework to tackle lung parenchyma segmentation and lung nodule detection simultaneously. The detection is performed on the basis of segmentation results in an anchor-free manner. Two simple and shallow ConvNets are used to predict the center position and size of the nodules on the segmentation feature map.

CPM-Net (Song et al. 2020) is a one-stage anchor-free detector consisting of a U-shaped 3D encoder-decoder backbone with two RPN heads that predict three maps in each head: classification map, size map, and offset map. The core of CPM-Net is an innovative training sample selection strategy called center point matching (CPM). This strategy calculates the distance between ground truths and sample points, taking the nearest K points as positive samples. Building upon CPM-Net, Luo et al. proposed SCPM-Net (Luo et al. 2022), which draws inspiration from the observation that nodules often have spherical or ellipsoidal morphology. A sphere-based nodule representation is proposed for training sample selection and defining regression terms, along with the introduction of a sphere-based loss function to locate nodules.

2.1.2.6 Different modeling methods Apart from the typical object detection frameworks and approaches, S4ND (Khosravan and Bagci 2018) modeled the nodule detection task as a cell-wise classification problem. It maps every 8 CT slices of size 512×512 to a probability map of size $8 \times 16 \times 16$, which determines whether there is a lung nodule in each 32×32 patch. This modeling method is simple and efficient, allowing for rapid inferring and low computational burden. However, it cannot provide a precise bounding box for each detected nodule.

In addition to the works based on supervised learning mentioned above, we find an interesting work (Yang et al. 2021a) that leverages additional electronic medical records (EMR) to provide weakly-supervised information. Specifically, Yang et al. pretrained a 3D-FPN detector on the LUNA16 and TianChi datasets and used weak labels, such as the number of nodules k and the slice index i where the nodules are located, from the patient's EMR to improve the robustness of the model. The proposed approach involves

Table 2 The performance and publication of lung nodule detectors

Paper	Publication	Dataset	0.125	0.25	0.5	1	2	4	8	CPM
Liao et al. (2019)	TNNLS, 2019	DSB17	–	–	–	–	–	–	–	85.6
NoduleNet (Tang et al. 2019c)	MICCAI, 2019	LIDC-IDRI	70.9	78.3	85.7	90.0	94.3	95.5	96.3	87.3
GA-SSD (Ma et al. 2019)	MIDL, 2019	LIDC-IDRI	–	–	–	–	–	–	–	86.3
		Private	–	–	–	–	–	–	–	73.3
13DR-Net (Harsono et al. 2022)	JKSUCIS, 2020	LIDC-IDRI	72.3	83.8	88.7	91.1	92.8	93.4	94.8	88.1
DeepSEED (Li and Fan 2020)	ISBI, 2020	LIDC-IDRI	73.9	80.3	85.8	88.8	90.7	91.6	92.0	86.2
		LUNA16	60.0	67.4	75.1	82.4	85.0	85.3	85.9	77.3
Ding et al. (2017)	MICCAI, 2017	LUNA16	–	–	–	–	–	–	–	89.3
Xie et al. (2019)	PR, 2018	LUNA16	73.4	74.4	76.3	79.6	82.4	83.2	83.4	79.0
Wang et al. (2018b)	MICCAI, 2018	LUNA16	67.6	77.6	87.9	94.9	95.8	95.8	95.8	87.8
S4ND (Khostravan and Bagci 2018)	MICCAI, 2018	LUNA16	–	–	–	–	–	–	–	89.7
Sun et al. (2018)	KSEM, 2018	LUNA16	73.4	74.4	76.3	79.6	82.4	83.2	83.4	79.0
DeepLung (Zhu et al. 2018)	WACV, 2018	LUNA16	–	–	–	–	–	–	–	84.2
Huang et al. (2019)	ISBI, 2019	LUNA16	–	–	–	91.4	–	94.6	–	88.0
3DFPN-HS ² (Liu et al. 2019b)	MICCAI, 2019	LUNA16	90.4	91.4	93.3	95.7	97.1	97.1	97.1	95.2
CPM-Net (Song et al. 2020)	MICCAI, 2020	LUNA16	72.3	83.8	88.7	91.1	92.8	93.4	94.8	88.1
LENS (Yan et al. 2020)	TMI, 2020	LUNA16	–	–	–	–	–	–	–	89.8
Liu et al. (2021b)	JBHI, 2021	LUNA16	–	–	–	–	–	–	–	93.9
TSND (Zheng et al. 2022)	Diagnostics, 2022	LUNA16	77.1	84.9	90.5	94.0	95.7	96.0	96.0	90.6
SCPM-Net (Luo et al. 2022)	MIA, 2022	LUNA16	74.3	82.9	88.9	92.2	93.9	95.8	96.4	89.2
DS-CMSF (Zhou et al. 2022)	JBHI, 2022	LUNA16	74.2	84.1	90.0	92.5	94.4	95.5	96.0	90.0
OSAF (Huang et al. 2022)	CMPB, 2022	LUNA16	78.8	85.2	89.0	92.8	95.4	96.1	96.2	90.5
Han et al. (2022)	CMPB, 2022	LUNA16	69.7	78.9	84.3	89.1	92.0	94.0	94.7	86.1
MSANet (Guo et al. 2021)	JBHI, 2022	LUNA16	66.8	76.4	81.9	86.1	90.8	91.2	91.4	83.5
Zhao et al. (2023)	ESWA, 2023	LUNA16	65.6	75.4	83.3	91.7	95.1	97.0	97.7	–
IR-Unet++ (Lin et al. 2023)	MBEC, 2023	LUNA16	–	–	–	90.1	–	94.8	95.8	87.4
SANet (Mei et al. 2021)	TPAMI, 2021	PN9	38.1	45.1	54.5	64.5	75.3	83.9	90.0	64.5

Table 2 (continued)

Paper	Publication	Dataset	0.125	0.25	0.5	1	2	4	8	CPM
LSSANet (Xu et al. 2022)	MICCAI, 2022	PN9	51.6	51.6	58.2	66.9	77.3	85.4	89.9	68.7
Tang et al. (2019b)	ISBI, 2019	TianChi	43.1	59.5	65.1	74.0	80.0	85.5	91.1	71.2
Yang et al. (2021a)	MICCAI, 2021	LUNA16	–	–	–	–	–	–	–	89.0
		TianChi	–	–	–	–	–	–	–	63.5
		Private	–	–	–	–	–	–	–	61.4
DeepLN (Xu et al. 2020)	KBS, 2020	Private	–	–	–	–	–	–	–	68.5

^a[0.125, 0.25, ..., 8] represents the number of FP predictions

Table 3 Publicly available mammographic datasets

Dataset	Modality	Cases	Images	View	Availability
DDSM (Heath et al. 1998)	SFM	2620	10,480	MLO, CC	Open access
INBreast (Moreira et al. 2012)	FFDM	115	410	MLO, CC	Need request
BCDR ^a	FFDM	1734	7315	MLO, CC	Open access
OPTIMAM (Halling-Brown et al. 2020)	DM	–	> 2 M	MLO, CC	Need request

<https://bcdre.eu/>

reclassifying the nodule candidates generated by the 3D-FPN detector. It selects the top k candidates with the highest predicted probabilities, along with the candidates with the highest predicted probabilities that are located on slice i , and treats them as positive samples.

Lung nodule detection, as the most extensively researched and developed area in tiny lesion detection, has witnessed significant advancements and has presented a relatively clear development path. The performance of lung nodule detectors we discussed in this section on their respective datasets is summarized in Table 2.

2.2 Breast masses

Breast cancer is the most prevalent cancer among women (Sung et al. 2021). A breast mass refers to an abnormal growth or tumor in the breast tissue, which can be as small as a few millimeters. Breast masses can be either benign or malignant. Benign masses typically have a spherical or oval shape, with well-defined and smooth borders. Malignant masses may have irregular, jagged, or spiculated borders, and there may be a distinct transition zone between the mass and the surrounding normal tissue. Detecting breast masses can facilitate early detection and diagnosis of breast cancer, thereby enabling timely intervention for treatment.

2.2.1 Datasets

We collected four publicly available datasets from the papers we gathered. These datasets comprise mammograms, which are low-energy X-ray images that detect abnormalities within the breast. These mammograms can be categorized into three types according to imaging techniques: DM (digital mammogram), FFDM (full-field digital mammogram), and SFM (screen film mammogram). Typically, two perspective images are captured for each breast, resulting in two distinct views: the cranio-caudal (CC) view and the medio-lateral oblique (MLO) view. The CC mammogram is taken from above, while the MLO mammogram is taken from the side. We summarized these datasets in Table 3 and provided separate introductions for each.

2.2.1.1 DDSM The digital database of screening mammography (DDSM, Heath et al. 1998) has a collection of 2620 scanned film-screen mammography of MLO and CC views. The dataset was published in 1996 and has contour points annotation of RoIs. It includes a well-annotated and labeled subset known as the curated breast imaging subset of the DDSM

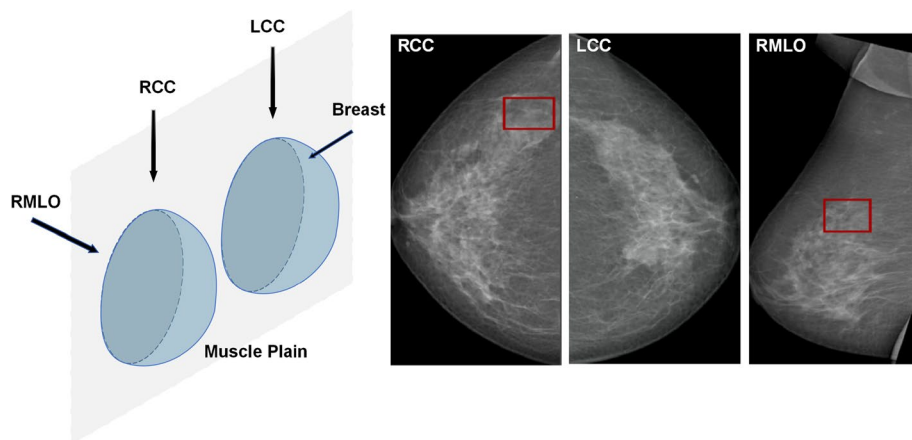


Fig. 12 An illustration of the various breast views

(CBIS-DDSM). This subset provides information about bounding boxes and segmentation mask for RoI, as well as detailed pathological data, including breast mass type, tumor grade, and stage.

2.2.1.2 INBreast The INBreast (Moreira et al. 2012) database consists of 410 screening mammograms obtained from 115 cases. Among these cases, 90 have MLO and CC views for both breasts, while the remaining 25 cases have these views available for only one breast. The annotations include information about abnormality types and the contour of masses.

2.2.1.3 BCDR The Breast Cancer Digital Repository (BCDR) includes 1734 cases of FFDs contributed by various institutions. The images are accessible in both the CC view and the MLO view. The dataset offers accurate lesion locations, BI-RADS density annotations, precise mass coordinates, and detailed segmentation outlines. Furthermore, auxiliary patient data, including prior surgery, lesion characteristics, and biopsy status, is also provided.

2.2.1.4 OPTIMAM The OPTIMAM (Halling-Brown et al. 2020) dataset is available upon request from the authors. The OPTIMAM database is a large collection of 2,889,312 images, and provides MLO and CC views and bounding box annotations of RoIs. OPTIMAM contains data from the first OPTIMAM1 project initiated in 2008, as well as the subsequent OPTIMAM2 project launched in 2013. OPTIMAM2 is one of the few public datasets that include annotated 3D digital breast tomosynthesis imaging.

2.2.2 Detectors

2.2.2.1 Single-view In 2015, Dhungel et al. were among the pioneers who applied DL for breast mass detection in Dhungel et al. (2015). They proposed a framework comprised of multiple components in sequence to detect mass candidates and gradually reduce FP predictions. Specifically, they employed a combination of a multi-scale deep belief network and a gaussian mixture model to generate mass candidates. Then, a cascade of R-CNN was employed to coarsely identify and locate masses. To further reduce FPs, knowledge-based

texture and morphological features were extracted and fed into a cascade of random forest classifiers. Finally, the finely masses prediction was provided through the post-processing of connected component analysis (CCA). In 2017, the authors made several improvements to this work in Dhungel et al. (2017). An additional hypothesis refinement stage was introduced to minimize redundant predictions, and CCA was replaced with NMS.

Moreover, there have been other studies that either adopt existing detectors or make minor improvements to them. For example, Faster R-CNN-like detectors were used in Akselrod-Ballin et al. (2016), Ribli et al. (2018), Cao et al. (2019) and Agarwal et al. (2020). RetinaNet was utilized in Jung et al. (2018). FSAF (Zhu et al. 2019) was applied in BMassDNet (Cao et al. 2021). Various versions of YOLO were employed in Al-Masni et al. (2017, 2018), Al-Antari et al. (2018) and YOLO-LOGO (Su et al. 2022).

2.2.2.2 Cross-/multi-view As shown in Fig. 12, in the clinical setting, the use of multi-view breast imaging can greatly aid decision-making process of radiologists. To mimic this practice, cross-/multi-view models have been proposed. Generally, three distinct combinations are employed: the bilateral perspective (contrasting views of both breasts) like Liu et al. (2019c), the ipsilateral perspective (top and side views of the same-side breast) as demonstrated in studies such as Ma et al. (2021), Liu et al. (2020), Yan et al. (2021), AlGhamdi and Abdel-Mottaleb (2021), and the three-view perspective (top and side views of the same-side breast, along with a top/contrasting view of the opposite-side breast) as shown in Yang et al. (2020a, 2021c) and Liu et al. (2021c). These studies adhere to the siamese network (Koch et al. 2015)-like pattern.

Both breasts of the same patient demonstrate a high level of symmetry. The corresponding regions in the image from the opposite side can be utilized for comparative verification. Liu et al. (2019c) proposed a model called contrasted bilateral network (CBN). Bilateral perspective images (RCC and LCC, where R and L denote right and left breast) are fed into CBN, and two modules are developed to leverage bilateral information: the distortion-insensitive comparison module (DICM) for handling non-rigid distortion and the logic-guided bilateral module (LGBM) for contrasting bilateral images. The entire pipeline can be seamlessly integrated into the Mask R-CNN framework for end-to-end training.

Ipsilateral perspective images consist of two views: the CC view and the MLO view. To leverage this relevant and complementary information, Ma et al. introduced a two-branch Faster R-CNN, named CVR-RCNN (Ma et al. 2021), to extract RoI features from both views. Additionally, they proposed cross-view relation networks to establish a grid connection between the RoI features. This facilitates the interaction of visual and geometric information between the two views, enabling simultaneous detection on both view branches. On the other hand, some studies introduced additional stages for lesion matching and FPRE based on the detection results to improve its accuracy. Yan et al. (2021) employed shared weight YOLOs to generate RoIs, and proposed a matching siamese network trained with contrastive loss for mass matching and re-identification. Liu et al. (2020) proposed the bipartite graph convolutional network (BGN). BGN constructs bipartite graph by mapping spatial visual features, where each node represents a relatively consistent region in the breasts, and the edges model both geometric constraints and appearance similarities between nodes. Graph convolution-based bipartite graph reasoning can enhance feature representation of masses. In DV-DCNN (AlGhamdi and Abdel-Mottaleb 2021), AlGhamdi et al. adopted RetinaNet as the detector and proposed

a 2D dual-view ConvNet for judging whether a pair of masses match. Non-matching masses are directly eliminated to reduce the FP rate of detection.

Radiologists routinely employ both bilateral and ipsilateral analysis in the diagnosis workflow. Building on this prior, Yang et al. proposed the MommiNet (Yang et al. 2020a), which is the first DL architecture for tri-view based breast mass detection. MommiNet comprises two primary sub-networks: BiDualNet and IpsiDualNet, which perform integrated analyses for ipsilateral and bilateral assessments, respectively. Concretely, the RCC view is designated as the main view, while the RMLO view and LCC view serve as auxiliary views, each paired with the main view as the input of the corresponding sub-network. BiDualNet predicts the segmentation results of the aligned bilateral views, while IpsiDualNet outputs the detection results based on nipple localization. The probability maps of the two sub-networks are integrated using a fusion module to generate the final prediction. In the following year, the authors upgraded the design of the sub-networks and fusion module, proposing MommiNet v2 (Yang et al. 2021c). Inspired by the MommiNet series, Liu et al. introduced AG-RCNN (Liu et al. 2021c), which is an update to their previous work (Liu et al. 2020) specifically for tri-view analysis. They incorporated contralateral analysis, resulting in a three-branch graph neural network with a similar overall structure to Yang et al. (2021c).

The unique dual-/tri-view modality of breast masses detection provides opportunities for designing accurate detectors. The performance and datasets of the included studies are shown in Table 4.

2.3 Thyroid nodules

Thyroid is an important endocrine organ in the human body. In the past 30 years, the incidence of thyroid-related diseases, especially thyroid cancer, has increased rapidly by 240% (Zhu et al. 2009). Early detection and screening are important ways to curb this trend. The formation of thyroid nodules is a typical early symptom of thyroid cancer. Clinical studies have shown that early diagnosis of thyroid nodules can effectively reduce the incidence and mortality rate of thyroid cancer (Bethesda 2018).

2.3.1 Datasets

As depicted in Fig. 13, many thyroid nodules exhibit indistinct boundaries. Ultrasound examination serves as a widely adopted method for rapid and non-invasive thyroid examination. The studies we reviewed primarily utilized private datasets, along with the inclusion of the public dataset DDTI (Pedraza et al. 2015). The details of these datasets are shown in Table 5.

2.3.2 Detectors

Taking into account the significant variation in the scale of thyroid nodules, the collected studies carefully configured anchors and performed detection on multi-scale feature maps. For instance, Abdolali et al. (2020) and configured the anchors by k-means clustering and genetic algorithm on their dataset to accommodate thyroid nodules, and utilized a 2D Mask R-CNN as the detector. Ma et al. also applied the k-means and proposed a variant of YOLOv3 called DMRF (Ma et al. 2020) as the detector.

Table 4 The performance and publication of breast masses detectors

Paper	Publication	Dataset	ACC	REC	TPR@FPi	AUC	mAP
Akselrod-Ballin et al. (2016)	DLMIA, 2016	Private	–	–	–	72.0	–
Al-Masni et al. (2017)	EMBC, 2017	DDSM	–	96.3	–	–	–
CBN (Liu et al. 2019c)	MICCAI, 2017	DDSM	–	–	69.1@0.50 78.8@1.00 85.6@2.00 87.7@3.00	–	–
Al-Masni et al. (2018)	CMPB, 2018	DDSM	99.7	–	99.2@0.22	–	–
BGN (Liu et al. 2020)	CVPR, 2020	DDSM	–	–	79.5@0.50 86.6@1.00 91.8@2.00 92.5@3.00	–	–
MommiNet (Yang et al. 2020a)	MICCAI, 2020	DDSM	–	–	80.0@0.50 85.0@1.00 89.0@2.00	–	–
MommiNet-v2 (Yang et al. 2021c)	MIA, 2021	DDSM	–	–	83.0@0.50 85.0@1.00 90.0@2.00	–	–
AGN (Liu et al. 2021c)	TPAMI, 2021	DDSM	–	–	82.0@0.50 89.0@1.00 92.1@2.00 93.8@3.00	–	–
CVR-RCNN (Ma et al. 2021)	ICPR, 2021	DDSM	–	–	85.0@1.20 88.0@1.90 92.0@4.40	–	–
Dhungel et al. (2015)	DICTA, 2015	INbreast	–	–	96.0@1.20 87.0@0.80	–	–
Dhungel et al. (2017)	MIA, 2017	INbreast	–	–	90.0@1.00	–	–
FrCN (Al-Antari et al. 2018)	IJMI, 2018	INbreast	99.0	–	–	–	–
Ribli et al. (2018)	Sci. Rep., 2018	INbreast	–	–	90.0@0.30	–	–
Jung et al. (2018)	PLoS ONE, 2018	INbreast	–	–	97.0@3.00 94.0@1.30 86.0@0.50	–	–
DeepLIM (Cao et al. 2019)	ICCV, 2017	CBIS-DDSM INbreast BCDR	–	97.7	–	–	80.5
Agarwal et al. (2020)	CIBM, 2019	CBIS-DDSM OPTIMAM	–	–	91.0@1.69 93.0@0.78	–	–
CMCNet (Yan et al. 2021)	MIA, 2021	CBIS-DDSM INbreast	–	–	96.0@0.23	94.8	–
DV-DCNN (AlGhamdi and Abdel-Mottaleb 2021)	CMPB, 2021	CBIS-DDSM INbreast	97.5 96.0	95.0 94.0	–	98.0 97.0	–
BMassDNet (Cao et al. 2021)	CMPB, 2021	DDSM INbreast	–	–	93.0@0.50 94.3@0.60	–	–
YOLO-LOGO (Su et al. 2022)	CMPB, 2022	CBIS-DDSM INbreast	–	95.7 92.5	–	–	65.0 61.4

^aTPR@FPi represents the TPR when predicting i FPs

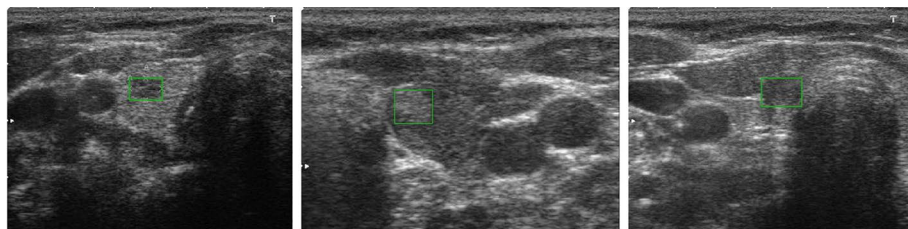


Fig. 13 Thyroid nodules in DDTI

Table 5 The involved datasets of thyroid nodules

Paper	Modality	Patients	Images	Dataset
M-SSD (Song et al. 2018)	Ultrasound, image	299	347	DDTI
	Ultrasound, image	1580	4309	Private
Liu et al. (2019a)	Ultrasound, image	4655	8140	Private
Abdolali et al. (2020)	Ultrasound, image	100	4102	Private
DMRF-CNN (Ma et al. 2020)	Ultrasound, image	240	699	Private
Yao et al. (2020)	Ultrasound, image	300	2459	Private
Wu et al. (2021)	Ultrasound, video	43	142	Private
FDnet (Song et al. 2022)	Ultrasound, image	299	347	DDTI
	Ultrasound, image	—	377	Private
Yu et al. (2023)	Ultrasound, image	1357	1921	Private

Since echoes can generate excessive noise in ultrasound images, and given that heterogeneous thyroid nodules with different components may resemble the background, several studies have proposed or utilized diverse feature fusion mechanisms to enhance the representation of these nodules. Song et al. proposed the MC-CNN (Song et al. 2018), a 2D SSD-based detector. They introduced nodule prior guided layers designed to extract RoI features of high-confidence anchors. These layers are attached to the VGG-16 backbone to improve the identification and localization of nodules. Liu et al. (2019a) and Yao et al. (2020) presented detectors based on the Faster R-CNN framework, both employing the FPN neck to hierarchically integrate features from different levels. In another work (Song et al. 2022), Song et al. not only introduced FPN to the Faster R-CNN but also designed a nodule segmentation branch with pseudo-label supervision behind FPN, which enhanced the features of each layer in FPN thus improving the detection accuracy. Inspired by the success of PANet (Liu et al. 2018), which incorporates a novel path aggregation mechanism, Yu et al. (2023) incorporated the PAN module, along with the CA module (Hou et al. 2021) and SPP module (He et al. 2015), into the CSPDarkNet (Redmon and Farhadi 2018; Wang et al. 2020) to construct the backbone and neck of the detector. This process integrates contextual information and provides a robust representation from different regions and scales.

Table 6 The performance and publication of thyroid nodule detectors

Paper	Publication	Dataset	mAP40	mAP50	mAP	ACC	REC	SPEC
M-SSD (Song et al. 2018)	JBHI, 2018	Private	–	–	97.0	98.2	98.3	98.0
		DDTI	–	–	92.1	94.1	96.2	
Liu et al. (2019a)	MIA, 2019	Private	–	–	95.0	92.8	96.4	78.0
Abdolali et al. (2020)	CBM, 2020	Private	–	–	82.0	–	79.0	–
DMRF-CNN (Ma et al. 2020)	CIN, 2020	Private	–	–	88.9	95.2	97.4	–
Yao et al. (2020)	FBB, 2020	Private	98.0	–	–	–	98.2	–
Wu et al. (2021)	JBHI, 2021	Private	–	88.2	61.0	–	–	–
FDnet (Song et al. 2022)	App. In., 2022	DDTI	–	75.8	48.6	–	–	–
		Private	–	92.5	61.8	–	–	–
Yu et al. (2023)	CBM, 2023	Private	–	89.0	50.1	–	85.1	–

Differing from the above image-based studies, Wu et al. (2021) proposed a detector that can directly process ultrasound videos, eliminating the need for manual key frame selection and making it more suitable for clinical practice. The detector is highly similar to the one in Yu et al. (2023). Notably, they proposed the cache track algorithm as a replacement for traditional NMS post-processing. This algorithm utilizes Kalman filter and Hungarian algorithm to correlate nodule predictions in adjacent images, imitating the diagnostic process of radiologists, to look forward and backward for correcting the FPs and FNs detection.

The performance of the collected models is shown in Table 6.

2.4 Diseased lymph nodes

Lymph nodes (LNs) are critical in the progression of cancer as they serve as sites for cancer spread (Nathanson 2003). When cancer cells break away from the primary tumor, they can travel through the lymphatic system and get trapped in nearby LNs and grow to form new tumors, leading to metastatic LNs. To establish optimal treatment staging, professionals rely on observing LNs from medical images and documenting their locations. An essential aspect of staging involves identifying metastatic LNs accurately, determining their position and size. However, given the significant individual variations, detecting LNs presents additional complexity (Torabi et al. 2004). Therefore, the development of an automated process is of importance to assist clinical practice in achieving accurate cancer staging.

2.4.1 Datasets

Different from the previously discussed lesions that are located in specific areas, LNs are a type of immune tissue that is distributed throughout the body. Figure 14 shows a LN example in a CT slice. Generally, LN datasets consist of either CT scans or MRI images. CT scans offer high-resolution 3D images and are relatively faster and more cost-effective. MRI, typically T2WI (T2-weighted imaging) and DWI (diffusion-weighted imaging), can reveal the presence of fluid content and pathological tissue changes, and provide functional information to help assess the pathological status of the LNs. LN detection studies we collected primarily relied on private datasets. The detailed information is shown in Table 7.

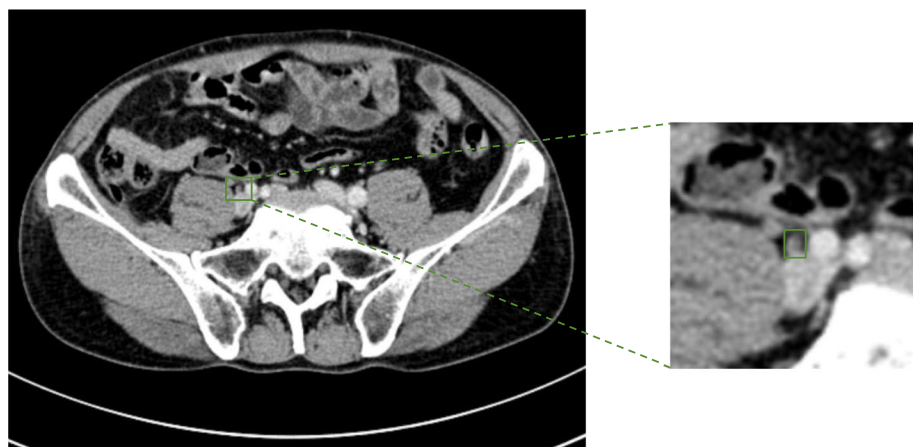


Fig. 14 A pelvic lymph node in CT image

Table 7 The datasets of lymph node involved

Paper	Modality	Patients	LNs	Dataset
Bouget et al. (2019)	CT	—	300	MeDataset ^a
Zhao et al. (2020)	T2WI + DWI	293	6724	private
Mathai et al. (2021, 2023)	T2WI	376	520	private
Wang et al. (2021)	CT	236	6880	private
Wang et al. (2022)	T2WI	584	821	private
Liu et al. (2022)	CT	401	800	private
Mathai et al. (2022)	T2WI	421	—	private
FAOT-Net (Zhang et al. 2023)	CT	236	6880	private

www.folk.ntnu.no/dnbouget/Mediastinal_Dataset/

2.4.2 Detectors

Mathai et al. conducted a series of studies on their dataset, including (Mathai et al. 2021, 2022, 2023). Mathai et al. (2021, 2023) ensembled several state-of-the-art models on COCO (Lin et al. 2014), such as FCOS (Tian et al. 2019), FoveaBox (Kong et al. 2020), VFNet (Zhang et al. 2021), etc, to detect abdominal LNs. On the other hand, Mathai et al. (2022) pioneered the use of detection transformer (DETR Carion et al. 2020).

Considering that a CT scan usually covers a larger area than the target region, Wang et al. (2021) proposed a two-step approach to detect abdominopelvic LNs: first localizing the abdominopelvic region and then detecting LNs within this region. To accomplish this, they employed a 3D ResNet-18 model to locate the starting frame and ending frame of the region. For the LN detection, they employed LeakyNet (Liao et al. 2019) as the detector. Following this work, Zhang et al. first proposed a 3D 1.5-stage detection framework called FAOT-Net (Zhang et al. 2023). FAOT-Net parallelizes the traditional two-stage process of predicting candidates and then performing FPR, enabling both processes to

Table 8 The performance and publication of lymph node detectors

Paper	Publication	TPR@FPi	PREC	mAP	CPM
Bouget et al. (2019)	IJCARS, 2019	74.2@9.00	–	–	–
Zhao et al. (2020)	EBioMedicine, 2020	80.0@8.60	73.5	–	–
Mathai et al. (2021)	MLMI, 2021	91.7@4.00	–	71.8	–
Wang et al. (2021)	IJIS, 2021	88.6@15.00	–	–	47.5
Wang et al. (2022)	MIA, 2021	73.9@5.85	–	47.8	–
Liu et al. (2022)	CMPB, 2022	65.7@/	59.7	45.4	–
Mathai et al. (2022)	SPIE MI, 2022	91.7@4.00	–	65.4	–
Mathai et al. (2023)	IJCARS, 2023	85.2@8.00	–	52.3	–
FAOT-Net (Zhang et al. 2023)	TMI, 2023	91.7@16.00	–	–	52.8

^aTPR@FPi represents the TPR when predicting i FPs

be supervised by a single loss function for consistent network optimization. In addition, Zhang et al. adopted the data-driven concept to redesign the anchor configuration strategy and positive–negative sample selection strategy. These together improved the FROC score for abdominopelvic lymph node detection by 5.3 points.

The use of R-CNN detectors remains prevalent. For instance, Liu et al. (2022) employed a 2D Faster R-CNN-based detector for axillary LN detection on 2D slices. Bouget et al. (2019) detected LNs on each CT slice separately, using a hybrid of 2D U-Net and 2D Mask R-CNN, and the detections were merged in the post-processing to form 3D bounding boxes.

Clinical prior is beneficial for algorithm development. Wang et al. (2022) introduced a 2D Mask R-CNN-based detector that utilizes RECIST bookmarks, commonly employed in the daily workflow of radiologists, to contour bounding boxes. Additionally, they proposed the global–local attention module to extract discriminative features and the multi-task uncertainty loss to weigh the multi-task losses. On the other hand, as T2WI and DWI are considered the most important sequences for nodal identification in clinical (Heijnen et al. 2013), Zhao et al. (2020) utilized various combinations of T2WI and high b-value DWI as 2D inputs for a 2D Mask R-CNN. Their model demonstrated good performance across external datasets, confirming its generalizability and the complementary nature of T2WI and DWI images.

As LNs are typically situated in the intricate and similar tissue background, it is challenging to detect them with high sensitivity while maintaining an acceptable number of FPs. The performance of the collected methods is shown in Table 8.

3 Universal lesion detection

Detection of single lesions has achieved promising accuracy overall. However, focusing solely on a specific type of lesion does not align with clinical workflow and may not provide sufficient guidance and reference for doctors. There exist correlations among certain lesions. For example, most advanced tumors have the potential to metastasize to lymph nodes. Therefore, obtaining a comprehensive overview of relevant clinical findings is crucial for making accurate diagnoses.

3.1 Datasets

Compared to general computer vision tasks, medical image analysis faces a shortage of large-scale and well-annotated datasets, primarily due to the much higher cost of annotation. In 2018, Yan et al. proposed a brand new CT scans dataset named “DeepLesion (Yan et al. 2018a)” which contains 32,120 CT slices from 10,594 studies of 4427 patients and includes 22 types of 32,735 lesion annotations. DeepLesion encompasses not only extensively studied lesion types but also less commonly researched ones, such as renal masses and iliac sclerotic lesions. Notably, the annotations in DeepLesion are directly extracted from doctors’ daily workflow, significantly reducing the burden of annotation.

3.2 Detectors

3.2.1 Work of DeepLesion team

In the initial paper (Yan et al. 2018a) introducing the DeepLesion dataset, a straightforward 2D Faster R-CNN is used as the detector, achieving a sensitivity of 81.1% with an average of 5 FPs per image. Since then, Yan et al. have iterated the model, consistently enhancing the detection accuracy and conducting valuable explorations.

Later that year, they presented the 3D context enhanced (3DCE Yan et al. 2018b) model that increased sensitivity to 84.37% at 4 FPs per image. 3DCE is built upon 2D R-FCN (Dai et al. 2016) and takes combined M (2+1)D images as input for a shared VGG-16-like backbone to produce M feature maps. These maps are concatenated in the channel dimension to generate the 3D-context-enhanced feature map. This strategy of deconstructing and fusing 3D information has since become a widely adopted paradigm in subsequent studies.

In 2019, Yan’s team proposed the universal lesion detector (ULDor Tang et al. 2019a) for simultaneous lesion segmentation and detection. ULDor is a 2D Mask R-CNN-based detector. Moreover, given that DeepLesion does not have fine-grained annotations for each lesion, they developed a lesion annotation network (LesaNet Yan et al. 2019a) to tag category labels for the lesions.

Detectors so far have been incapable of offering complete clinical assistance within a single model. To better fit the workflow of radiologists, Yan et al. proposed the multi-task universal lesion analysis network (MULAN, Yan et al. 2019b). MULAN is capable of simultaneously handling detection, segmentation and tagging, mimicking the doctors’ process of searching for lesions, characterizing and measuring them, and subsequently describing them in radiology reports. MULAN is a 2D Mask R-CNN-based model that also utilizes consecutive (2+1)D data as input. The main distinction from the original 2D Mask R-CNN is the incorporation of an additional tagging head, which assigns labels to lesions like LesaNet. It is interesting that the inclusion of the tagging branch shows a promotion to the detection, achieving an average sensitivity of 86.12% at 0.5 to 4 FPs. Moreover, the backbone of MULAN, a 2D DenseNet-121 (Huang et al. 2017) with the fourth dense block removed, has been widely adopted in subsequent studies.

Inspired by ACS (Yang et al. 2021b) to perform native 3D information extraction while utilizing pretrained weights on 2D natural image datasets, Yan’s team proposed P3DC (Cai et al. 2020b) (pseudo 3D convolution). P3DC is an anchor-free one-stage 3D lesion detector that extends the size of 2D convolution kernels by unsqueezing an additional dimension, thus creating 3D kernels and forming a pseudo 3D network. Moreover, they

incorporated the idea of representative points introduced in RepPoints (Yang et al. 2019), generating more precise bounding boxes by regressing a set of surface points.

In 2020 and 2021, Yan et al. proposed lesion-harvester (LH, Cai et al. 2020a) and lesion ensemble (LENS, Yan et al. 2020) models to find and label missing annotated lesions. LENS takes a step forward to universal lesion detection, involving multiple datasets, namely the universal lesion dataset DeepLesion, the lung nodule dataset LUNA16, the liver tumor dataset LiTS, and the lymph node dataset NIH-LN. While DeepLesion lacks fully annotations, the rest are considered to be fully annotated for a single type of lesion. LENS is a two-stage anchor-free detector following the classic paradigm of first extracting candidates in stage I, and then reducing FPs in stage II. However, what sets LENS apart is the inclusion of a separate RPN head for each dataset, referred to as the anchor-free proposal network (AFP), in order to adapt to different domains. Furthermore, LENS proposes several missing annotation mining strategies to collect possible lesions. First, the intra-patient lesion matching strategy matches the detection results of multiple CT scans taken at different time for individual patients. Moreover, the cross-dataset lesion mining strategy ignores the FP predictions of the single-type AFPs, while treating the new annotations obtained through box propagation and intra-patient lesion matching as true lesions during the retraining process. LENS can effectively detect unmarked lesions. Notably, it achieved an average sensitivity of 0.898 on the LUNA16 dataset, outperforming most of models designed specifically for this dataset.

3.2.2 Other work

Yan et al. have made numerous valuable contributions to ULD. Meanwhile, there are other noteworthy works from different teams. For instance, instead of simply concatenating M feature maps, Tao et al. (2019) introduced a contextual and spatial attention module to select relevant slices and mine discriminative regions. Zhang et al. proposed the 3D fusion module (3DFM Zhang et al. 2019) to learn and combine 3D information. Li et al. (2023) used morphology operations to achieve multi-scale inputs and designed a BiFPN-like (Tan et al. 2020) neck to enhance features. They also employed a multi-point regression strategy to increase the number of positive anchors. Inspired by 3DCE, PAC-Net (Xu et al. 2023) incorporated the SPP (He et al. 2015) block and Faster R-CNN structure, and proposed position attention guided connection (PAC) blocks for feature enhancement. On the other hand, several studies have followed the trend of incorporating multi-scale detection and introducing attention mechanisms to enhance feature representation. These include the works of Shao et al. (2019) (MSB), Wang et al. (2019, VA Mask R-CNN), Zlocha et al. (2019), and Liu et al. (2021a) (MLANet).

ElixirNet (Jiang et al. 2020) is the first work we have observed that utilizes neural architecture search (NAS, Elsken et al. 2019) in medical imaging detection. ElixirNet is developed on the foundation of a 2D Mask R-CNN and utilizes (2+1)D data as input. Its truncated RPN designs an adaptive filtering mechanism to reduce the number of region proposals by half, which speeds up inferring while providing potential lesions with higher confidence. Furthermore, the NAS strategy is adopted to construct ALB and ALBr modules, aiming to better explore the feature representations of the potential lesions. Additionally, a relation transfer module is proposed to address the issue of scarce labeled lesions.

Considering the 3D nature of CT scans, concatenating (2+1)D image features to obtain 3D features, as seen in 3DCE, MULAN, and others, might be the suboptimal choice. Hence, Yang et al. introduced ACS (Yang et al. 2021b) (axial-coronal-sagittal) convolution

to perform native 3D information extraction while utilizing pretrained weights on 2D datasets. In ACS, 3D kernels are divided into three parts based on the channel, and convoluted separately on the axial, coronal and sagittal views of the 3D representations. Except for ACS (Yang et al. 2021b) and P3DC (Cai et al. 2020b) mentioned above, other studies on convolution operators between 2020 and 2021 include MP3D (Zhang et al. 2020), AlignShift (Yang et al. 2020b), A3D (Yang et al. 2021d). MP3D utilized depth-wise separable convolution to break down 3D convolution into depth-wise convolution and point-wise convolution. AlignShift aimed to bridge the performance gap between thin- and thick-slice volumes by introducing virtual slices. A3D incorporated dense linear connections within the slice dimension for each channel to weight slices and fuse 3D context from different 2D slices. Compared with traditional 2D and 3D convolution kernels. These operators strive to reduce computation and memory costs of the 3D models while make more efficient use of 3D CT images.

In addition to these, MVP-Net (Li et al. 2019) verified the effectiveness of integrating multi-window views. Three commonly inspected windows, with window levels and window widths of [50, 449], [− 505, 1980], and [446, 1960] respectively. Experiments have demonstrated that the inclusion of multiple views can increase sensitivity by approximately

Table 9 The performance and publication of detectors in DeepLesion

Paper	Publication	0.125	0.25	0.5	1	2	4	8	16
3DCE (Yan et al. 2018b)	MICCAI, 2018	–	–	62.5	73.4	80.7	85.7	89.1	91.1
GPN (Li 2019)	MLMI, 2019	–	–	45.0	56.0	65.0	72.0	75.0	77.0
Tao et al. (2019)	MICCAI, 2019	–	–	71.4	78.5	84.0	87.6	90.2	91.4
Zhang et al. (2019)	MICCAI, 2019	41.1	53.8	66.3	76.3	82.9	87.0	89.8	91.7
MSB (Shao et al. 2019)	MICCAI, 2019	–	–	67.0	76.8	83.7	87.0	89.0	92.0
Zlocha et al. (2019)	MICCAI, 2019	–	–	72.2	80.1	86.4	90.8	94.1	96.3
VA Mask R-CNN (Wang et al. 2019)	MICCAI, 2019	–	–	69.1	77.9	83.8	–	–	–
MVP-Net (Li et al. 2019)	MICCAI, 2019	–	–	73.8	81.8	87.6	91.3	–	–
ULDor (Tang et al. 2019a)	ISBI, 2019	–	–	52.9	64.8	74.8	84.4	87.2	91.8
MULAN (Yan et al. 2019b)	MICCAI, 2019	–	–	76.1	83.7	88.8	92.3	94.7	95.6
P3DC (Cai et al. 2020b)	MICCAI, 2020	–	55.1	64.5	69.2	77.6	–	–	–
MP3D (Zhang et al. 2020)	MICCAI, 2020	–	–	79.6	85.3	89.6	92.5	–	–
AlignShift (Yang et al. 2020b)	MICCAI, 2020	–	–	79.4	85.5	90.1	93.3	95.2	96.7
ElixirNet (Jiang et al. 2020)	AAAI, 2020	–	–	65.6	75.8	83.7	88.9	91.4	–
Li et al. (2020)	MICCAI, 2020	–	–	73.3	81.2	86.8	90.7	–	–
Li et al. (2021)	MICCAI, 2021	–	–	79.2	85.7	89.8	92.7	–	–
ACS (Yang et al. 2021b)	JBHI, 2021	–	–	78.4	85.4	90.1	93.2	95.2	96.8
A3D (Yang et al. 2021d)	MICCAI, 2021	–	–	80.3	86.7	91.3	94.1	96.2	97.3
MLANet (Liu et al. 2021a)	EAAI, 2021	–	–	–	77.1	83.0	88.3	–	–
RECIST-Net (Xie et al. 2021)	ISBI, 2021	–	–	76.1	83.7	89.6	92.5	–	–
DSA-ULD (Sheoran et al. 2022)	ISBI, 2022	–	–	78.3	84.5	89.0	92.4	–	–
Li et al. (2023)	ESA, 2023	–	–	76.8	84.1	84.8	86.9	89.9	–
PAC-Net (Xu et al. 2023)	FBB, 2023	49.3	62.9	73.2	81.1	86.9	91.4	92.2	–
LH (Cai et al. 2020a)	TMI, 2020	25.5	38.8	47.1	54.3	59.5	64.1	71.7	–
LENS (Yan et al. 2020)	TMI, 2020	23.7	31.6	40.3	50.0	59.6	69.5	78.0	–

^a[0.125, 0.25, \ldots, 8] represents the number of FP predictions

3 points. Following the windows setting of MVP-Net, Xie et al. proposed RECIST-Net (Xie et al. 2021), an anchor-free detector inspired by ExtremeNet (Zhou et al. 2019), which detects lesions by locating the center point and four extreme points of RECIST bookmarks. DSA-ULD (Sheoran et al. 2022), derived from FCOS (Tian et al. 2019), expanded the number of windows to five.

Lesion boundaries can often be approximated using ellipses. Building upon this prior knowledge, Li reconstructed the conventional bounding box regression problem as a bounding ellipse regression, which takes inspiration from face detection. He developed the gaussian proposal network (GPN, Li 2019), rather than using region proposal network (RPN) as usual, to locate lesions. Except for GPN, Li et al. reconstructed regression targets by transforming bounding boxes into bounding maps in Li et al. (2020, 2021). Instead of improving data, operators, or models, these different approaches to modeling detection tasks provide us a fresh perspective.

Compared to traditional SLD, ULD-related work to some extent transcends the category of fixed tasks of detection, moving towards a more comprehensive integration in an attempt to provide more comprehensive and integrated clinical services. It has formed a prototype of a multimodal big medical model with certain ‘image-report’ characteristics. The detection performance of the models involved in this section are shown in Table 9. Note that the indicators of LH and LENS correspond to the missing annotated lesion mining, they should not be directly compared with other detectors.

4 Discussion

Throughout the field of lesion detection, numerous insightful methods and models have been proposed, enabling the state-of-the-art models to achieve sensitivity comparable to that of professional doctors in internal validation. However, several challenges and issues are present.

Data poses the primary challenge. Annotating medical images requires expert knowledge. The usual approach is to involve multiple annotators. However, it is hard to sustain and expand due to the substantial cost of medical annotators. Furthermore, annotators typically specialize in specific subdivisions. These facts restrict the dataset’s capacity and diversity. Lesion detection datasets remain considerably smaller and are insufficiently labeled compared to natural image datasets. Data will remain one of the major obstacles in the foreseeable future. Exploring broader and deeper cooperation with medical institutions and minimizing the annotation burden should be the primary considerations. The use of pretrained AI models to achieve a semi-automated annotation process should be valued.

Models usually have insufficient generalizability. Factors such as noise, motion artifacts, image blur, and variations in scanning parameters introduce diversity and damage the performance of models, making them struggle to achieve the same performance in external validation. Robust models necessitate robust feature representations. Existing studies have primarily built upon supervised learning, leveraging transfer learning, multi-task learning, or attention mechanisms to obtain improved representations. However, considering the scarcity of labeled data and the abundance of unlabeled data, it is beneficial to find appropriate agent tasks for training feature extractors through unsupervised or semi-supervised learning to achieve more robust feature representations.

Moving towards universality is the prevailing trend. In 2022, ChatGPT,⁴ a language model trained on massive datasets with trillions of parameters, sparked worldwide discussions on ‘big model’ and ‘foundation model’. The following year, several large language models for healthcare were also released, including Med-PaLM (Singhal et al. 2023) and MediSearch.⁵ These models provide services such as answering, diagnosing, making treatment recommendations, and more. In the field of medical imaging, MedSAM (Ma and Wang 2023) utilizes over 200,000 masks from 11 different modalities to fine-tune the SAM (segment anything model Kirillov et al. 2023), a vision transformer-based large vision model, enhancing its suitability for medical image segmentation. In today’s booming era of foundation models, the ULD is undoubtedly a crucial advancement in whole-body lesion detection. Furthermore, thanks to the multimodal capability of ULD (such as MULAN), ULD might become a key step towards a foundational model with the ability of handling ‘image-report-treatment’. The medical AI community still has a long way to go.

The ULD expands the scope of tiny lesion detection. However, there are still some critical downstream tasks based on lesion detection that need to be addressed to better meet clinical needs, such as lesion tracking and learning feedback. Firstly, doctors often need to identify and track the same lesion in a patient across time sequential images to evaluate its progression. The detection-based lesion tracking models are expected to assist in this process. Secondly, doctors expect AI to provide feedback. If AI can explain its reasoning and provide evidence for each detected lesion, it has the potential to offer new insights and guidance into clinical practice and drive advancements in medical imaging. Automated lesion detection holds immense potential. We sincerely anticipate that it will continually optimize the workflow of medical imaging and improve the clinical situation.

5 Conclusion

This paper analyzes the evolution of DL-based tiny lesion detection from single to universal. We provide an overview and discussion of representative models and techniques to gain a comprehensive understanding of their relationships, highlight the specific issues they address, and present corresponding solutions. Our analysis covers the detection of lung nodules, breast masses, thyroid nodules, lymph nodes, as well as universal lesions. In addition, we summarize the challenges encountered, offer potential solutions, and identify future research directions. The objective of this review is to enhance readers’ understanding of DL models in detecting tiny lesions, and their potential in improving efficacy, accuracy, and consistency in various lesion detection tasks. We hope that these findings will inspire future research endeavors.

Author contributions Conception/writing: Yi Zhang. Writing/paper collecting: Yiji Mao, Xuanyu Lu, Xingyu Zou, and Xinyang Li. Check/validation: Jiayue Li. Medical knowledge support: Hao Huang.

Funding The work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0100201, and in part by the Major Science and Technology Project from the Science & Technology Department of Sichuan Province under Grant 2022ZDZX0023.

Data availability No datasets were generated or analysed during the current study.

⁴ <https://openai.com/blog/chatgpt/>.

⁵ <https://medisearch.io/>.

Declarations

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abdolali F, Kapur J, Jaremko JL, Noga M, Hareendranathan AR, Punithakumar K (2020) Automated thyroid nodule detection from ultrasound imaging using deep convolutional neural networks. *Comput Biol Med* 122:103871
- Agarwal R, Diaz O, Yap MH, Lladó X, Marti R (2020) Deep learning for mass detection in full field digital mammograms. *Comput Biol Med* 121:103774
- Akselrod-Ballin A, Karlinsky L, Alpert S, Hasoul S, Ben-Ari R, Barkan E (2016) A region based convolutional network for tumor detection and classification in breast mammography. In: Deep learning and data labeling for medical applications: first international workshop, LABELS 2016, and 2nd international workshop, DLMIA 2016, held in conjunction with MICCAI 2016, Athens, Greece, 21 Oct 2016, proceedings, pp 197–205. Springer, Singapore
- Al-Antari MA, Al-Masni MA, Choi M-T, Han S-M, Kim T-S (2018) A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. *Int J Med Informatics* 117:44–54
- AlGhamdi M, Abdel-Mottaleb M (2021) DV-DCNN: dual-view deep convolutional neural network for matching detected masses in mammograms. *Comput Methods Programs Biomed* 207:106152
- Al-Masni MA, Al-antari MA, Park J, Gi G, Kim T-Y, Rivera P, Valarezo E, Han S-M, Kim T-S (2017) Detection and classification of the breast abnormalities in digital mammograms via regional convolutional neural network. In: 2017 39th Annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp 1230–1233
- Al-Masni MA, Al-Antari MA, Park J-M, Gi G, Kim T-Y, Rivera P, Valarezo E, Choi M-T, Han S-M, Kim T-S (2018) Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system. *Comput Methods Programs Biomed* 157:85–94
- Armato SG III, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA et al (2011) The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys* 38(2):915–931
- Bethesda M (2018) Seer cancer stat facts thyroid cancer. National Cancer Institute, Bethesda. Accessed 10 May 2021
- Bouget D, Jørgensen A, Kiss G, Leira HO, Langø T (2019) Semantic segmentation and detection of mediastinal lymph nodes and anatomical structures in CT data for lung cancer staging. *Int J Comput Assist Radiol Surg* 14:977–986
- Cao Z, Yang Z, Zhuo X, Lin R-S, Wu S, Huang L, Han M, Zhang Y, Ma J (2019) Deeplima: Deep learning based lesion identification in mammograms. In: Proceedings of the IEEE/CVF international conference on computer vision workshops
- Cai J, Harrison AP, Zheng Y, Yan K, Huo Y, Xiao J, Yang L, Lu L (2020a) Lesion-Harvester: iteratively mining unlabeled lesions and hard-negative examples at scale. *IEEE Trans Med Imaging* 40(1):59–70
- Cai J, Yan K, Cheng C-T, Xiao J, Liao C-H, Lu L, Harrison AP (2020b) Deep volumetric universal lesion detection using light-weight pseudo 3D convolution and surface point regression. In: Medical image computing and computer assisted intervention—MICCAI 2020: 23rd International conference, Lima, Peru, 4–8 Oct 2020, proceedings, Part IV. Springer, Cham, pp 3–13
- Cao H, Pu S, Tan W, Tong J (2021) Breast mass detection in digital mammography based on anchor-free architecture. *Comput Methods Programs Biomed* 205:106033

- Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision. Springer, Cham, pp 213–229
- Carreira J, Zisserman A (2017) Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308
- Dai J, Li Y, He K, Sun J (2016) R-FCN: object detection via region-based fully convolutional networks. In: Advances in neural information processing systems, vol 29
- Dhungel N, Carneiro G, Bradley AP (2017) A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Med Image Anal* 37:114–128
- Dhungel N, Carneiro G, Bradley AP (2015) Automated mass detection in mammograms using cascaded deep learning and random forests. In: 2015 International conference on digital image computing: techniques and applications (DICTA). IEEE, pp 1–8
- Ding J, Li A, Hu Z, Wang L (2017) Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In: Medical image computing and computer assisted intervention—MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, 11–13 Sept 2017, proceedings, Part III. Springer, Cham, pp 559–567
- Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q (2019) Centernet: keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6569–6578
- Elsken T, Metzen JH, Hutter F (2019) Neural architecture search: a survey. *J Mach Learn Res* 20(1):1997–2017
- George J, Skaria S, Varun V et al (2018) Using YOLO based deep learning network for real time detection and localization of lung nodules from low dose CT scans. In: Medical imaging 2018: computer-aided diagnosis, vol 10575. SPIE, pp 347–355
- Girshick R (2015) Fast R-CNN. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
- Guo Z, Zhao L, Yuan J, Yu H (2021) MSANet: multiscale aggregation network integrating spatial and channel information for lung nodule detection. *IEEE J Biomed Health Inf* 26(6):2547–2558
- Halling-Brown MD, Warren LM, Ward D, Lewis E, Mackenzie A, Wallis MG, Wilkinson LS, Given-Wilson RM, McAviney R, Young KC (2020) Optimam mammography image database: a large-scale resource of mammography images and clinical data. *Radiol Artif Intell* 3(1):200103
- Han Y, Qi H, Wang L, Chen C, Miao J, Xu H, Wang Z, Guo Z, Xu Q, Lin Q et al (2022) Pulmonary nodules detection assistant platform: an effective computer aided system for early pulmonary nodules detection in physical examination. *Comput Methods Prog Biomed* 217:106680
- Harsono IW, Liawatimena S, Cenggoro TW (2022) Lung nodule detection and classification from thorax CT-scan using retinanet with transfer learning. *J King Saud Univ Comput Inf Sci* 34(3):567–577
- He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 37(9):1904–1916
- He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
- Heath M, Bowyer K, Kopans D, Kegelmeyer Jr P, Moore R, Chang K, Munishkumaran S (1998) Current status of the digital database for screening mammography. In: Digital mammography: Nijmegen. Springer, Dordrecht, pp 457–460
- Heijnen LA, Lambregts DM, Mondal D, Martens MH, Riedl RG, Beets GL, Beets-Tan RG (2013) Diffusion-weighted mr imaging in primary rectal cancer staging demonstrates but does not characterise lymph nodes. *Eur Radiol* 23:3354–3360
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
- Hou Q, Zhou D, Feng J (2021) Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13713–13722
- Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
- Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
- Huang X, Sun W, Tseng T-LB, Li C, Qian W (2019) Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic CT scans using deep convolutional neural networks. *Comput Med Imaging Graph* 74:25–36
- Huang Y-S, Chou P-R, Chen H-M, Chang Y-C, Chang R-F (2022) One-stage pulmonary nodule detection using 3-D DCNN with feature fusion and attention mechanism in CT image. *Comput Methods Programs Biomed* 220:106786

- Jaeger PF, Kohl SA, Bickelhaupt S, Isensee F, Kuder TA, Schlemmer H-P, Maier-Hein KH (2020) Retina U-Net: embarrassingly simple exploitation of segmentation supervision for medical object detection. In: Machine learning for health workshop. PMLR, pp 171–183
- Jiang C, Wang S, Liang X, Xu H, Xiao N (2020) Elixirnet: relation-aware network architecture adaptation for medical lesion detection. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 11093–11100
- Jung H, Kim B, Lee I, Yoo M, Lee J, Ham S, Woo O, Kang J (2018) Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. *PLoS ONE* 13(9):0203355
- Khosravan N, Bagci U (2018) S4ND: single-shot single-scale lung nodule detection. In: Medical image computing and computer assisted intervention—MICCAI 2018: 21st international conference, Granada, Spain, 16–20 Sept 2018, proceedings, Part II. Springer, Cham, pp 794–802
- Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y et al (2023) Segment anything. *arXiv preprint. arXiv:2304.02643*
- Koch G, Zemel R, Salakhutdinov R et al (2015) Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop, vol 2. Lille, France
- Kong T, Sun F, Liu H, Jiang Y, Li L, Shi J (2020) FoveaBox: beyond anchor-based object detector. *IEEE Trans Image Process* 29:7389–7398
- Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, vol 25
- Law H, Deng J (2018) Cornernet: Detecting objects as paired keypoints. In: Proceedings of the european conference on computer vision (ECCV), pp 734–750
- Li Y (2019) Detecting lesion bounding ellipses with gaussian proposal networks. In: Machine learning in medical imaging: 10th international workshop, MLMI 2019, held in conjunction with MICCAI 2019, Shenzhen, China, 13 Oct 2019, proceedings. Springer, Cham, pp 337–344
- Li Y, Fan Y (2020) DeepSpeed: 3D squeeze-and-excitation encoder-decoder convolutional neural networks for pulmonary nodule detection. In: 2020 IEEE 17th International symposium on biomedical imaging (ISBI). IEEE, pp 1866–1869
- Li Z, Zhang S, Zhang J, Huang K, Wang Y, Yu Y (2019) Mvp-net: multi-view fpn with position-aware attention for deep universal lesion detection. In: Medical image computing and computer assisted intervention—MICCAI 2019: 22nd international conference, Shenzhen, China, 13–17 Oct 2019, proceedings, Part VI. Springer, Cham, pp 13–21
- Li H, Han H, Zhou SK (2020) Bounding maps for universal lesion detection. In: International conference on medical image computing and computer-assisted intervention, pp. 417–428. Springer
- Li H, Chen L, Han H, Chi Y, Zhou SK (2021) Conditional training with bounding map for universal lesion detection. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 141–152
- Li H, Rong H, Sheng VS, Song Y, Qiu C, Han K, Liu Z (2023) Cascaded multi-point regression network for high-quality generic lesion detection. *Expert Syst Appl* 214:119141
- Liao F, Liang M, Li Z, Hu X, Song S (2019) Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. *IEEE Trans Neural Netw Learn Syst* 30(11):3484–3495
- Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: Common objects in context. In: Computer vision—ECCV 2014: 13th European conference, Zurich, Switzerland, 6–12 Sept 2014, proceedings, Part V. Springer, Cham, pp 740–755
- Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017a) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125
- Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017b) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
- Lin J, She Q, Chen Y (2023) Pulmonary nodule detection based on IR-Unet++. *Med Biol Eng Comput* 61(2):485–495
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: Single shot multibox detector. In: Computer vision—ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, proceedings, Part I 14, pp 21–37. Springer
- Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8759–8768
- Liu T, Guo Q, Lian C, Ren X, Liang S, Yu J, Niu L, Sun W, Shen D (2019a) Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks. *Med Image Anal* 58:101555

- Liu J, Cao L, Akin O, Tian Y (2019b) 3dfpn-hs² 2: 3D feature pyramid network based high sensitivity and specificity pulmonary nodule detection. In: Medical image computing and computer assisted intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, proceedings, Part VI 22. Springer, Cham, pp 513–521
- Liu Y, Zhou Z, Zhang S, Luo L, Zhang Q, Zhang F, Li X, Wang Y, Yu Y (2019c) From unilateral to bilateral learning: Detecting mammogram masses with contrasted bilateral network. In: Medical image computing and computer assisted intervention—MICCAI 2019: 22nd international conference, Shenzhen, China, 13–17 Oct 2019, proceedings, Part VI 22. Springer, Cham, pp 477–485
- Liu Y, Zhang F, Zhang Q, Wang S, Wang Y, Yu Y (2020) Cross-view correspondence reasoning based on bipartite graph convolutional network for mammogram mass detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3812–3822
- Liu Z, Xie X, Song Y, Zhang Y, Liu X, Zhang J, Sheng VS (2021a) MLANet: multi-layer anchor-free network for generic lesion detection. *Eng Appl Artif Intell* 102:104255
- Liu W, Liu X, Li H, Li M, Zhao X, Zhu Z (2021b) Integrating lung parenchyma segmentation and nodule detection with deep multi-task learning. *IEEE J Biomed Health Inform* 25(8):3073–3081
- Liu Y, Zhang F, Chen C, Wang S, Wang Y, Yu Y (2021c) Act like a radiologist: towards reliable multi-view correspondence reasoning for mammogram mass detection. *IEEE Trans Pattern Anal Mach Intell* 44(10):5947–5961
- Liu Z, Huang D, Yang C, Shu J, Li J, Qin N (2022) Efficient axillary lymph node detection via two-stage spatial-information-fusion-based cnn. *Comput Methods Programs Biomed* 223:106953
- Luo X, Song T, Wang G, Chen J, Chen Y, Li K, Metaxas DN, Zhang S (2022) SCPM-Net: an anchor-free 3D lung nodule detection network using sphere representation and center points matching. *Med Image Anal* 75:102287
- Ma J, Li X, Li H, Menze BH, Liang S, Zhang R, Zheng W-S (2019) Group-attention single-shot detector (ga-ssd): finding pulmonary nodules in large-scale ct images. In: International conference on medical imaging with deep learning. PMLR, pp 358–369
- Ma J, Duan S, Zhang Y, Wang J, Wang Z, Li R, Li Y, Zhang L, Ma H (2020) Efficient deep learning architecture for detection and recognition of thyroid nodules. *Computational Intelligence and Neuroscience*. 2020
- Ma J, Li X, Li H, Wang R, Menze B, Zheng W-S (2021) Cross-view relation networks for mammogram mass detection. In: 2020 25th international conference on pattern recognition (ICPR). IEEE, pp 8632–8638
- Mathai TS, Lee S, Elton DC, Shen TC, Peng Y, Lu Z, Summers RM (2021) Detection of lymph nodes in t2 mri using neural network ensembles. In: Machine learning in medical imaging: 12th international workshop, MLMI 2021, held in conjunction with MICCAI 2021, Strasbourg, France, 27 Sept 2021, proceedings. Springer, Cham, pp 682–691
- Mathai TS, Lee S, Elton DC, Shen TC, Peng Y, Lu Z, Summers RM (2022) Lymph node detection in t2 mri with transformers. In: Medical imaging 2022: computer-aided diagnosis, vol 12033. SPIE, pp 855–859
- Mathai TS, Lee S, Shen TC, Lu Z, Summers RM (2023) Universal lymph node detection in t2 mri using neural networks. *Int J Comput Assist Radiol Surg* 18(2):313–318
- Ma J, Wang B (2023) Segment anything in medical images. *arXiv preprint. arXiv:2304.12306*
- Mei J, Cheng M-M, Xu G, Wan L-R, Zhang H (2021) Sanet: A slice-aware network for pulmonary nodule detection. *IEEE Trans Pattern Anal Mach Intell* 44(8):4374–4387
- Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS (2012) Inbreast: toward a full-field digital mammographic database. *Acad Radiol* 19(2):236–248
- Nathanson SD (2003) Insights into the mechanisms of lymph node metastasis. *Cancer* 98(2):413–423
- Pedraza L, Vargas C, Narváez F, Durán O, Muñoz E, Romero E (2015) An open access thyroid ultrasound image database. In: 10th International symposium on medical information processing and analysis, vol 9287. SPIE, pp 188–193
- Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement. *arXiv preprint. arXiv:1804.02767*
- Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
- Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, vol 28
- Ribli D, Horváth A, Unger Z, Pollner P, Csabai I (2018) Detecting and classifying lesions in mammograms with deep learning. *Sci Rep* 8(1):4165
- Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention—MICCAI 2015: 18th

- international conference, Munich, Germany, 5–9 Oct 2015, proceedings, Part III. Springer, Cham, pp 234–241
- Setio AAA, Traverso A, De Bel T, Berens MS, Van Den Bogaard C, Cerello P, Chen H, Dou Q, Fantacci ME, Geurts B et al (2017) Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Med Image Anal* 42:1–13
- Shao Q, Gong L, Ma K, Liu H, Zheng Y (2019) Attentive ct lesion detection using deep pyramid inference with multi-scale booster. In: Medical image computing and computer assisted intervention—MICCAI 2019: 22nd international conference, Shenzhen, China, 13–17 Oct 2019, proceedings, Part VI. Springer, Cham, pp 301–309
- Sheoran M, Dani M, Sharma M, Vig L (2022) An efficient anchor-free universal lesion detection in CT-scans. In: 2022 IEEE 19th international symposium on biomedical imaging (ISBI). IEEE, pp 1–4
- Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pföhl S et al (2023) Large language models encode clinical knowledge. *Nature* 620(7972):172–180
- Song W, Li S, Liu J, Qin H, Zhang B, Zhang S, Hao A (2018) Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition. *IEEE J Biomed Health Inf* 23(3):1215–1224
- Song T, Chen J, Luo X, Huang Y, Liu X, Huang N, Chen Y, Ye Z, Sheng H, Zhang S et al (2020) CPM-Net: a 3D center-points matching network for pulmonary nodule detection in CT scans. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 550–559
- Song R, Zhu C, Zhang L, Zhang T, Luo Y, Liu J, Yang J (2022) Dual-branch network via pseudo-label training for thyroid nodule detection in ultrasound image. *Appl Intell* 52(10):11738–11754
- Su Y, Liu Q, Xie W, Hu P (2022) Yolo-logo: a transformer-based yolo segmentation model for breast mass detection and segmentation in digital mammograms. *Comput Methods Programs Biomed* 221:106903
- Sun N, Yang D, Fang S, Xie H (2018) Deep convolutional nets for pulmonary nodule detection and classification. In: Knowledge science, engineering and management: 11th international conference, KSEM 2018, Changchun, China, 17–19 Aug 2018, proceedings, Part II. Springer, pp 197–208
- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 71(3):209–249
- Tang Y-B, Yan K, Tang Y-X, Liu J, Xiao J, Summers RM (2019a) Uldor: a universal lesion detector for ct scans with pseudo masks and hard negative example mining. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). IEEE, pp 833–836
- Tang H, Liu X, Xie X (2019b) An end-to-end framework for integrated pulmonary nodule detection and false positive reduction. In: 2019 IEEE 16th International symposium on biomedical imaging (ISBI 2019). IEEE, pp 859–862
- Tang H, Zhang C, Xie X (2019c) NoduleNet: decoupled false positive reduction for pulmonary nodule detection and segmentation. In: Medical image computing and computer assisted intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, 13–17 Oct 2019, proceedings, Part VI. Springer, Cham, pp 266–274
- Tan M, Pang R, Le QV (2020) EfficientDet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10781–10790
- Tao Q, Ge Z, Cai J, Yin J, See S (2019) Improving deep lesion detection using 3D contextual and spatial attention. In: Medical image computing and computer assisted intervention—MICCAI 2019: 22nd international conference, Shenzhen, China, 13–17 Oct 2019, proceedings, Part VI. Springer, Cham, pp 185–193
- Tian Z, Shen C, Chen H, He T (2019) FCOS: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9627–9636
- Torabi M, Aquino SL, Harisinghani MG (2004) Current concepts in lymph node imaging. *J Nucl Med* 45(9):1509–1518
- Tsai C-H, Peng Y-S (2022) Multi-task lung nodule detection in chest radiographs with a dual head network. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 707–717
- Wang X, Girshick R, Gupta A, He K (2018a) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
- Wang B, Qi G, Tang S, Zhang L, Deng L, Zhang Y (2018b) Automated pulmonary nodule detection: high sensitivity with few candidates. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 759–767

- Wang X, Han S, Chen Y, Gao D, Vasconcelos N (2019) Volumetric attention for 3D medical image segmentation and detection. In: Medical image computing and computer assisted intervention—MICCAI 2019: 22nd international conference, Shenzhen, China, 13–17 Oct 2019, proceedings, Part VI. Springer, Cham, pp 175–184
- Wang C-Y, Liao H-YM, Wu Y-H, Chen P-Y, Hsieh J-W, Yeh I-H (2020) CSPNet: a new backbone that can enhance learning capability of CNN. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 390–391
- Wang H, Huang H, Wang J, Wei M, Yi Z, Wang Z, Zhang H (2021) An intelligent system of pelvic lymph node detection. *Int J Intell Syst* 36(8):4088–4116
- Wang S, Zhu Y, Lee S, Elton DC, Shen TC, Tang Y, Peng Y, Lu Z, Summers RM (2022) Global-local attention network with multi-task uncertainty loss for abnormal lymph node detection in MR images. *Med Image Anal* 77:102345
- Wu X, Tan G, Zhu N, Chen Z, Yang Y, Wen H, Li K (2021) CacheTrack-YOLO: Real-Time Detection and Tracking for thyroid nodules and surrounding tissues in ultrasound videos. *IEEE J Biomed Health Inform* 25(10):3812–3823
- Xie H, Yang D, Sun N, Chen Z, Zhang Y (2019) Automated pulmonary nodule detection in ct images using deep convolutional neural networks. *Pattern Recogn* 85:109–119
- Xie C, Cao S, Wei D, Zhou H, Ma K, Zhang X, Qian B, Wang L, Zheng Y (2021) RECIST-Net: lesion detection via grouping keypoints on recist-based annotation. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). IEEE, pp 921–924
- Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500
- Xu X, Wang C, Guo J, Yang L, Bai H, Li W, Yi Z (2020) Deepln: a framework for automatic lung nodule detection using multi-resolution CT screening images. *Knowl Based Syst* 189:105128
- Xu R, Luo Y, Du B, Kuang K, Yang J (2022) LSSANet: a long short slice-aware network for pulmonary nodule detection. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 664–674
- Xu Z, Li T, Liu Y, Zhan Y, Chen J, Lukasiewicz T (2023) PacNet: multi-pathway FPN with position attention guided connections and vertex distance iou for 3d medical image detection. *Front Bioeng Biotechnol* 11:1049555
- Yan K, Wang X, Lu L, Summers RM (2018a) Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *J Med Imag* 5(3):036501
- Yan K, Bagheri M, Summers RM (2018b) 3D context enhanced region-based convolutional neural network for end-to-end lesion detection. In: Medical image computing and computer assisted intervention—MICCAI 2018: 21st international conference, Granada, Spain, 16–20 Sept 2018, proceedings, Part I. Springer, Cham, pp 511–519
- Yan K, Peng Y, Sandfort V, Bagheri M, Lu Z, Summers RM (2019a) Holistic and comprehensive annotation of clinically significant findings on diverse CT images: learning from radiology reports and label ontology. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 8523–8532
- Yan K, Tang Y, Peng Y, Sandfort V, Bagheri M, Lu Z, Summers RM (2019b) MULAN: multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In: Medical image computing and computer assisted intervention—MICCAI 2019: 22nd international conference, Shenzhen, China, 13–17 Oct 2019, proceedings, Part VI. Springer, Cham, pp 194–202
- Yan K, Cai J, Zheng Y, Harrison AP, Jin D, Tang Y, Tang Y, Huang L, Xiao J, Lu L (2020) Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in CT. *IEEE Trans Med Imaging* 40(10):2759–2770
- Yan Y, Conze P-H, Lamard M, Quéllec G, Cochenier B, Coatrieux G (2021) Towards improved breast mass detection using dual-view mammogram matching. *Med Image Anal* 71:102083
- Yang Z, Liu S, Hu H, Wang L, Lin S (2019) RepPoints: point set representation for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9657–9666
- Yang Z, Cao Z, Zhang Y, Han M, Xiao J, Huang L, Wu S, Ma J, Chang P (2020a) MommiNet: mammographic multi-view mass identification networks. In: Medical image computing and computer assisted intervention—MICCAI 2020. Springer, Cham
- Yang J, He Y, Huang X, Xu J, Ye X, Tao G, Ni B (2020b) Alignshift: bridging the gap of imaging thickness in 3D anisotropic volumes. In: Medical image computing and computer assisted intervention—MICCAI 2020: 23rd international conference, Lima, Peru, 4–8 Oct 2020, proceedings, Part IV. Springer, Cham, pp 562–572

- Yang H-H, Wang F-E, Sun C, Huang K-C, Chen H-W, Chen Y, Chen H-C, Liao C-Y, Kao S-H, Wang Y-CF et al (2021a) Leveraging auxiliary information from EMR for weakly supervised pulmonary nodule detection. In: Medical image computing and computer assisted intervention—MICCAI 2021: 24th international conference, Strasbourg, France, 27 Sept–1 Oct 2021, proceedings, Part VII, pp. 251–261. Springer
- Yang J, Huang X, He Y, Xu J, Yang C, Xu G, Ni B (2021b) Reinventing 2D convolutions for 3D images. *IEEE J Biomed Health Inform* 25(8):3009–3018
- Yang Z, Cao Z, Zhang Y, Tang Y, Lin X, Ouyang R, Wu M, Han M, Xiao J, Huang L et al (2021c) Mommi-Net-v2: mammographic multi-view mass identification networks. *Med Image Anal* 73:102204
- Yang J, He Y, Kuang K, Lin Z, Pfister H, Ni B (2021d) Asymmetric 3D context fusion for universal lesion detection. In: Medical image computing and computer assisted intervention—MICCAI 2021: 24th international conference, Strasbourg, France, 27 Sept–1 Oct 2021, proceedings, Part V. Springer, Cham, pp. 571–580
- Yao S, Yan J, Wu M, Yang X, Zhang W, Lu H, Qian B (2020) Texture synthesis based thyroid nodule detection from medical ultrasound images: interpreting and suppressing the adversarial effect of in-place manual annotation. *Front Bioeng Biotechnol* 8:599
- Yu Z, Liu S, Liu P, Liu Y (2023) Automatic detection and diagnosis of thyroid ultrasound images based on attention mechanism. *Comput Biol Med* 155:106468
- Zhang Z, Zhou Y, Shen W, Fishman E, Yuille A (2019) Lesion detection by efficiently bridging 3D context. In: Machine learning in medical imaging: 10th international workshop, MLMI 2019, held in conjunction with MICCAI 2019, Shenzhen, China, 13 Oct 2019, proceedings. Springer, Cham, pp 471–478
- Zhang S, Xu J, Chen Y-C, Ma J, Li Z, Wang Y, Yu Y (2020) Revisiting 3D context modeling with supervised pre-training for universal lesion detection in CT slices. In: Medical image computing and computer assisted intervention—MICCAI 2020: 23rd international conference, Lima, Peru, 4–8 Oct 2020, proceedings, Part IV. Springer, Cham, pp 542–551
- Zhang H, Wang Y, Dayoub F, Sunderhauf N (2021) VarifocalNet: an IoU-aware dense object detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8514–8523
- Zhang Y, Li J, Li X, Xie M, Islam MT, Zhang H (2023) FAOT-Net: a 1.5-stage framework for 3D pelvic lymph node detection with online candidate tuning. *IEEE Trans Med Imaging* 43(3):1180–1190
- Zhao X, Xie P, Wang M, Li W, Pickhardt PJ, Xia W, Xiong F, Zhang R, Xie Y, Jian J et al (2020) Deep learning-based fully automated detection and segmentation of lymph nodes on multiparametric-MRI for rectal cancer: a multicentre study. *EBioMedicine* 56:102780
- Zhao D, Liu Y, Yin H, Wang Z (2023) An attentive and adaptive 3D CNN for automatic pulmonary nodule detection in CT image. *Expert Syst App* 211:118672
- Zheng S, Kong S, Huang Z, Pan L, Zeng T, Zheng B, Yang M, Liu Z (2022) A lower false positive pulmonary nodule detection approach for early lung cancer screening. *Diagnostics* 12(11):2660
- Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J (2018) UNet++: a nested U-Net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, 20 Sept 2018, proceedings. Springer, Cham, pp 3–11
- Zhou X, Zhuo J, Krahenbuhl P (2019) Bottom-up object detection by grouping extreme and center points. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 850–859
- Zhou Z, Gou F, Tan Y, Wu J (2022) A cascaded multi-stage framework for automatic detection and segmentation of pulmonary nodules in developing countries. *IEEE J Biomed Health Inform* 26(11):5619–5630
- Zhu C, Zheng T, Kilfoy BA, Han X, Ma S, Ba Y, Bai Y, Wang R, Zhu Y, Zhang Y (2009) A birth cohort analysis of the incidence of papillary thyroid cancer in the united states, 1973–2004. *Thyroid* 19(10):1061–1066
- Zhu W, Liu C, Fan W, Xie X (2018) Deeplung: Deep 3D dual path nets for automated pulmonary nodule detection and classification. In: 2018 IEEE Winter conference on applications of computer vision (WACV). IEEE, pp 673–681
- Zhu C, He Y, Savvides M (2019) Feature selective anchor-free module for single-shot object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 840–849
- Zlocha M, Dou Q, Glocker B (2019) Improving retinanet for ct lesion detection with dense masks from weak recist labels. In: Medical image computing and computer assisted intervention—MICCAI 2019: 22nd international conference, Shenzhen, China, 13–17 Oct 2019, proceedings, Part VI. Springer, Cham, pp 402–410

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yi Zhang¹ · Yiji Mao¹ · Xuanyu Lu¹ · Xingyu Zou¹ · Hao Huang² · Xinyang Li¹ ·
Jiayue Li³ · Haixian Zhang¹

✉ Haixian Zhang
zhanghaixian@scu.edu.cn

Yi Zhang
zhangyi_magic@scu.stu.edu.cn

Yiji Mao
maoyiji@scu.stu.edu.cn

Xuanyu Lu
luxuanyu@scu.stu.edu.cn

Xingyu Zou
zouxingyu@scu.stu.edu.cn

Hao Huang
huanghao96@scu.stu.edu.cn

Xinyang Li
lixinyang@scu.stu.edu.cn

Jiayue Li
laylali34@outlook.com

¹ Machine Intelligence Lab, School of Computer Science and Technology, Sichuan University, Chengdu 610065, Sichuan, China

² Colorectal Cancer Center, West China Hospital, Sichuan University, Chengdu 610041, Sichuan, China

³ Google LLC, 1600 Amphitheatre Pkwy, Mountain View, Mountain View, CA 95133, USA