electronics

Article

MWSR-YLCA: Improved YOLOv7 Embedded with Attention Mechanism for Nasopharyngeal Carcinoma Detection from MR Images

Huixin Wu [1], Xin Zhao [1], Guanghui Han [1,2,*], Haojiang Li [3,*], Yuhao Kong [1] and Jiahui Li [1]

[1] School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, China; wuhuixin@ncwu.edu.cn (H.W.); zhaoxin@stu.ncwu.edu.cn (X.Z.)
[2] School of Biomedical Engineering, Sun Yat-sen University, Shenzhen 518107, China
[3] State Key Laboratory of Oncology in South China, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
* Correspondence: hanguanghui@ncwu.edu.cn (G.H.); lihaoj@sysucc.org.cn (H.L.)

Abstract: Nasopharyngeal carcinoma (NPC) is a malignant tumor, and early diagnosis and timely treatment are important for NPC patients. Accurate and reliable detection of NPC lesions in magnetic resonance (MR) images is very helpful for the disease diagnosis. However, recent deep learning methods need to be improved for NPC detection in MR images. Because NPC tumors are invasive and usually small in size, it is difficult to distinguish NPC tumors from the closely connected surrounding tissues in a huge and complex background. In this paper, we propose an automatic detection method, named MWSR-YLCA, to accurately detect NPC lesions in MR images. Specifically, we design two modules, the multi-window settings resampling (MWSR) module and an improved YOLOv7 embedded with a coordinate attention mechanism (YLCA) module, to detect NPC lesions more accurately. First, the MWSR generates a pseudo-color version of MR images based on a multi-window resampling method, which preserves richer information. Subsequently, the YLCA detects the NPC lesion areas more accurately by constructing a novel network based on an improved YOLOv7 framework embedded with the coordinate attention mechanism. The proposed method was validated on an MR image set of 800 NPC patients and obtained 80.1% mAP detection performance with only 4694 data samples. The experimental results show that the proposed MWSR-YLCA method can perform high-accuracy detection of NPC lesions and has superior performance.

Keywords: nasopharyngeal carcinoma; multi-window resampling; attention mechanism; object detection

Citation: Wu, H.; Zhao, X.; Han, G.; Li, H.; Kong, Y.; Li, J. MWSR-YLCA: Improved YOLOv7 Embedded with Attention Mechanism for Nasopharyngeal Carcinoma Detection from MR Images. *Electronics* **2023**, *12*, 1352. https://doi.org/10.3390/electronics12061352

Academic Editors: Sathishkumar V E and Malliga Subramanian

Received: 16 December 2022
Revised: 7 March 2023
Accepted: 7 March 2023
Published: 12 March 2023

1. Introduction

Nasopharyngeal carcinoma (NPC) is a malignant tumor that occurs in the nasopharyngeal site and lateral wall, and is endemic in southern China, North Africa, and Southeast Asia [1]. According to the data of the World Health Organization [2] in 2021, the number of new cases of NPC diagnosed globally reached 133,000. The incidence of NPC in China is higher than the average incidence rate in the world. New NPC cases in China account for about 50% of the world's total. Incidences of NPC is one of the highest of malignant tumors in China, and the incidence is highest in otolaryngology malignant tumors. Therefore, research on NPC needs to continue, and a new synergistic relationship between distant metastasis in patients with nasopharyngeal carcinoma has been discovered in the latest study [3]. Detailed examination of magnetic resonance (MR) imaging is necessary to accurately depict the primary tumor, and, as a routine clinical procedure for the diagnosis of NPC, preoperative MR is used to assess tumor progression. Most nasopharyngeal cancers are moderately sensitive to radiation therapy, and radiation therapy is the treatment of choice for NPC. Tumor detection and segmentation of MR images is an important step in computer-aided tumor diagnosis [4]. A reliable automatic detection model can

quickly detect tumor areas and effectively reduce the radiation therapy planning workload of radiologists.

At present, there are few studies on the detection of NPC. NPC tumors usually occupy a small volume in MR images, and the tissue background is closely connected to the tumor, with complex and variable border shapes that are difficult to distinguish and sometimes impossible to identify by the human eye, making the detection of NPC lesions more challenging. To address these issues, several segmentation methods for NPC detection have been proposed in previous work. For example, in 2015, Huang et al. [5] introduced a region-based NPC segmentation method, which used clustering and classification methods to segment nasopharyngeal carcinoma from MR images. In 2018, Mohammed et al. [6] proposed a new method for diagnosing NPCs from endoscopic images, which includes a trainable segmentation of NPC tissues, a genetic algorithm to select the best features, and a support vector machine for classifying NPCs, and the detection shows high accuracy. The disadvantage is that this method needs to display tumors on several incisions, and doctors need to draw separate ROIs on different tumor incisions to detect NPC segmentation in one patient, which is complex and requires a lot of time and money. In 2019, Zhao et al. [7] proposed a DL method, which used a deep convolutional neural network (DCNN) to achieve automatic NPC segmentation on 2D PET-CT images with dice similarity coefficient (DSC), sensitivity, and accuracy of 0.785, 0.764, 0.789, respectively. However, it only collected PET-CT images from 22 patients, with a small number of samples, and increased the complexity of image data. In 2020, Chen et al. [8] proposed a new multimodal MR fusion network (MMFNet) based on a multi-encoder and introduced a 3D-CBAM attention module to highlight information features. This method mainly uses different forms of MR images to complete accurate segmentation. The above methods [5–7] are complex and demanding in terms of data processing for NPC detection, increasing the complexity of NPC images (multimodality). Their experiments uses 256, 381, and 1100 images, respectively, and the data scale is small. Our experiment uses 4694 NPC MR images for lesion detection, and the results are more reliable. The existing NPC segmentation methods for MR images need to be improved in the aspects of data processing cost, data scale, and algorithm performance.

The difficulty of NPC detection is partly due to the small size of the NPC lesion area in proportion to the whole MR image, while the complex background occupies the major part. Moreover, the shape of nasopharyngeal carcinoma is diverse, the background and tumor boundary are blurred, and the lesion shape is complex and difficult to distinguish, which makes NPC detection very difficult. In 2022, Wang et al. [9] proposed a new network based on an improved Mask R-CNN framework using global-local attention to detect abnormal lymph nodes in MR images with good performance. Inspired by this literature, we introduced an attention mechanism [10] to enhance the feature representation of NPC lesion regions and weaken the influence of background regions.

The traditional detection method of NPC takes a long time, and the accuracy of the algorithm is affected by the way of image feature extraction [11,12]. It uses manual operation, which is costly and has a high tendency for errors. In contrast, deep neural networks have powerful automatic representation learning capabilities. As one of the main architectures of DL, the convolutional neural network (CNN) method provides superior performance for classification, segmentation, and detection tasks in digital pathological images (DPGA) [13]. Therefore, the CNN-based architecture is often used as a tool for faster and more accurate diagnosis by processing multimodal MRI images [14]. Deep neural networks have powerful automatic representation learning capabilities and have been widely used in the detection and segmentation tasks of medical images. Zhang at al. [15] developed a computer-aided detection method based on the deep learning model Faster R-CNN, which has the potential to detect brain metastases with high sensitivity and reasonable specificity. Elakkiya at al. [16] proposed a hybrid deep learning technology, which developed a small object detection generative adversarial network (SOD-GAN) based on RCNN to automatically detect and classify cervical precancerous lesions and malignant lesions according to deep features without any preliminary classification and segmentation

assistance. However, the R-CNN and Faster R-CNN methods only focus on the number of randomly selected and determined regions, and do not scan the entire input image, so they may miss key regions or concentrate on unimportant regions. These two cases will lead to error detection and classification results, while the YOLO object detector can scan the entire input image for classification and area detection. Salman at al. [17] developed an automatic tool for detection and diagnosis of prostate cancer based on the YOLO algorithm, and empirical results demonstrate that it is possible to develop high-performance prostate cancer diagnostic tools using the object detection method. These diagnostic tools can reduce inter-observer variation between pathologists and decrease time delay in the diagnostic phase. Salman at al. [17] also show that the YOLO algorithm can have good performance in cancer detection. In addition, there is no research on NPC detection in MR images using YOLO algorithm. In order to solve the above challenges of NPC detection, we improved the YOLO algorithm by integrating MWSR and YLCA modules, which can more accurately locate the object when performing lesion detection, and have higher detection performance and faster real-time detection of NPC lesions.

Deep neural network-based object detectors continue to evolve and are used in various applications. Object detectors accomplish both classification and localization by providing the location of the object as well as category labels and confidence scores, which are essential in high-impact real-world applications, and new methods are constantly being proposed. The CNN-based One-stage object detection OverFeat [18], YOLO [19], SSD [20], and RetinaNet [21] are improved step by step and are the basis for subsequent research in the object detection domain. In 2016, Redmon at al. [19] proposed the YOLO (You Only Look Once) algorithm, which treats object detection as a spatial location regression problem containing category information and forms a new paradigm for object detection. After continuous optimization and innovation, Wang at al [22] proposed YOLOv7 detector in 2022, and conducted the validation experiments on PASCALVOC and MS-COCO datasets. Empirical results demonstrate that YOLOv7 outperforms all known object detectors in the range of 5 FPS to 160 FPS in terms of speed and accuracy. YOLOv7 algorithm is the most advanced method for object detection and classification. We use the detector based on YOLOv7 algorithm to perform NPC detection. The main reason is that it determines the lesion area more accurately by analyzing the input features of the entire image. When using the image training detector in a specific field, it performs positioning and classification tasks more accurately than the previous algorithm, and has higher positioning and classification rates. Moreover, the algorithm detects real-time object faster than other algorithms [22].

Medical images are more complex and have greater variability than natural scene images. In recent years, convolutional neural networks (CNNs) have been successfully applied to automatic medical image detection, and the automatic detection of NPC in MR images has effectively reduced the doctor's workload in NPC diagnosis. In this paper, we propose an automatic method (MWSR-YLCA) for detection and diagnosis of NPC. Specifically, we design two modules in the MWSR-YLCA method, the multi-window settings resampling (MWSR) module and an improved YOLOv7 with an embedded coordinate attention mechanism (YLCA) module, to detect NPC lesions more accurately. First, the MWSR processes MR images of NPC through an image resampling method based on multi-window settings, which uses a windowing technique to fuse the optimal window width window position and nearby image information to enrich the amount of image information. Subsequently, the new YLCA network is constructed by embedding the fusion attention mechanism to enhance the feature representation of objects of interest for automatic detection and diagnosis of NPC. Due to the lack of public NPC detection data sets, we trained and evaluated our proposed model on our collected data sets, which include 26,000 MR images of 800 patients. By conducting extensive experiments using 4694 MR images containing lesion annotations, we evaluated the effectiveness of our proposed MWSR-YLCA and obtained high-accuracy NPC lesion detection performance. This paper main contributions are as follows.

(1) We use the multi-window setting based image resampling method (MWSR) to process NPC MR images. This method uses window technology to fuse image information in several windows (the optimal window and nearby windows), which reduces the information loss of the original image and enriches the image information for model input. The NPC detection performance using our method is improved compared to the detection performance using the original image, which provides a new way for medical MR images for NPC detection.
(2) We propose an NPC detection network YLCA for automatic detection and diagnosis of NPC, which builds a new network based on a YOLOv7 object detection network, embeds the fusion attention mechanism, and designs MP-CA Block to enhance the feature representation of objects of interest. Through extensive experimental evaluation, our detection network obtained the highest 80.2% mAP and 0.77 F1 compared to other comparison methods, proving that it is more effective for NPC MR image detection.

2. Method

The MWSR-YLCA method proposed in this paper consists of two main parts to jointly realize the detection of NPC lesions. The first part (MWSR) uses multi-window technology to resample the NPC MR image to obtain a three-channel (RGB) pseudo-color image for model training evaluation. The second part (YLCA) is based on the YOLOv7 [22] framework, embedding the coordinate attention (CA [23]) mechanism, constructing the attention convolution module MP-CA, obtaining the attention features, and fusing the attention features to construct the YLCA network, thereby improving the detection performance of the network.

2.1. Window Technique

The window technology in the field of medical images includes window width (WW) and window level (WL), which are used to select the range of CT values of interest. Because each tissue structure has a different range of CT values, when displaying a certain tissue structure, the suitable window width and window level for observing the tissue or lesion should be selected to obtain the best display effect. MR images are reconstructed analogue digital grey-scale images and therefore also have the characteristics to obtain the best display and perform various image post-processing using windowing techniques. However, unlike CT, the grey scale on MR images does not represent the density of soft tissues and lesions, but rather their MR signal intensity, reflecting the length of the relaxation time, and, therefore, the windowing technique for MR imaging does not have a fixed window width/level, which needs to be adjusted for each image. The DICOM image protocol specifies that medical images need to be stored as 16 bits (the actual number of bits used may be different), which indicates that the brightness of the pixel will be expressed in 2^{16} gray levels, and the role of window technology is to take out the grayscale value in a certain range of pixels in a 2^{16} grayscale image to display according to its gray level (usually 2^8), so as to display more image details.

Figure 1 shows the MR image window width and window level diagram. First, we set a range, and the gray value range of the observed tissue is listed separately, called the window. The gray value of a certain range is taken from the MR grayscale range and mapped to the gray image. The tissue whose gray value is higher than the window range is displayed as white; tissues below this window range appear black, then the size of this MR grayscale range is called window width WW, and the central value of this grayscale range is called window level WL.

2.2. YOLOv7

The convolutional neural network (CNN) is a major branch of neural networks and one of the main algorithms for deep learning in image applications [24,25]. It is a deep feedforward neural network with three characteristics of local connection, weight sharing, and down sampling, which can effectively reduce the complexity of the network and prevent the occurrence of

overfitting. The core feature of CNN is based on the convolution kernel, which is composed of several convolution layers, pooling layers, and fully connected layers. The convolution layer extracts different features of input image through convolution operation. The pooling layer reduces the feature dimension of data by partitioning the features. For the image, the main function of the pooling layer is to compress image features. The fully connected layer connects the extracted features to generate global features for image classification.

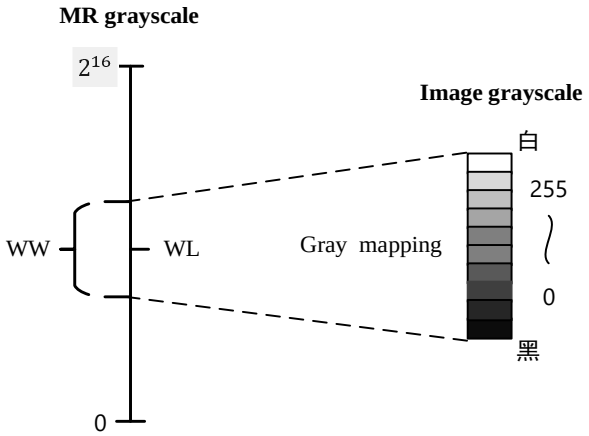


Figure 1. MR image window width and window level diagram.

The YOLOv7 [22] model was proposed in 2022 and validated on the COCO dataset to obtain better performance, standing out with faster speed and higher accuracy compared to the latest object detectors and attracting much attention. The general architecture of YOLOv7 consists of backbone, neck, and head. The entire network structure of 106 layers, of which the backbone layer is 51 and the head part 55, and consists mainly of modules such as CBS, MP, ELAN, and SPPCSPC. The structure of each module is shown in Figure 2. Compared to the previous YOLO model, YOLOv7 has been architecturally reformed using E-ELAN [22] and composite model scaling. It outperforms all real-time object detectors in terms of speed and accuracy, and it improves performance while reducing parameters by 40% and calculations by 50%.

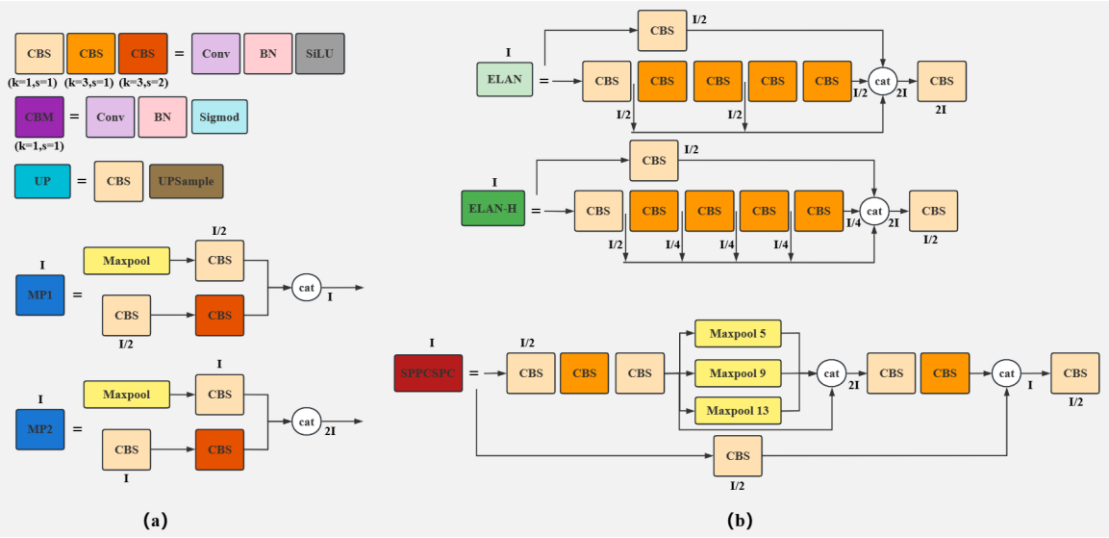


Figure 2. YOLOv7 main module diagram. (**a**) is the basic network module in YOLO. (**b**) is the improvement module of YOLOv7 compared to the previous YOLO series.

2.3. Attention Mechanism

The attention mechanism is derived from the study of human vision. Generally speaking, because humans have a limited capacity to process information, they selectively focus on the more important part of all information and ignore the rest. The attention mechanism is similar to the logic that humans use to look at pictures. When we look at a picture, we do not see the whole picture, but we focus our attention on the focal point of the picture. The core logic of an attention mechanism is focusing on the focal point instead of focusing on the whole. The attention mechanism has been widely used to achieve good performance in a variety of computer vision tasks, such as image classification, image segmentation, and object detection. The attention mechanism in neural networks is mainly implemented through the attention score. The attention score is a digital value between 0 and 1, and the sum of all scores under the attention mechanism is 1. Each attention score represents the attention weight assigned to the current item. Attention mechanisms can make the neural network ignore unimportant feature vectors and focus on calculating useful feature vectors. While eliminating the interference of unimportant features on the fitting results, the operation speed is improved. There are many types of attention mechanisms, such as channel attention [26], spatial attention [27], self-attention [28], mutual attention [29], coordinate attention [23], mixed attention, etc.

For mobile networks, the standardized attention mechanism SE (squeeze-and-excitation attention) effectively constructs the interdependence between channels by simply squeezing each two-dimensional feature map, which is significantly effective for improving the performance of the model. However, SE [26] attention only considers the importance of encoding inter-channel information while ignoring location information, which largely influences the generation of selective attention maps and is important for focusing on feature regions of interest. In this paper, experiments are conducted using the embedded coordinate attention CA [23], which splits the channel SE [26] into two parallel 1D feature encodings, and clusters the features separately in the two directions. The method embeds localization information in channel attention, enabling it to capture long-range correlations in one spatial direction while maintaining accurate location information in another, effectively integrating spatial coordinate information into the generated attention map. These graphs are applied to the input feature maps to enhance the representation of objects of interest by supplementing the feature map information, which is essential for locating object regions in computer vision tasks. The schematic diagram of the CA network structure is shown in Figure 3.

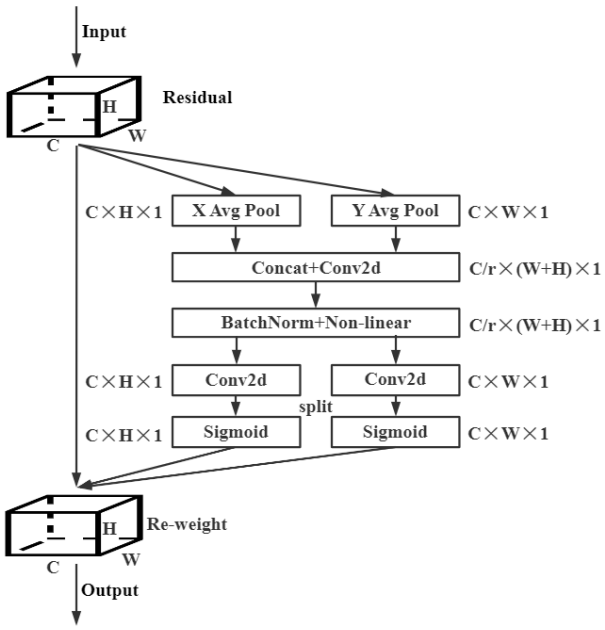


Figure 3. CA attention mechanism network structure diagram.

CA [23] divides the attention mechanism into two stages to encode channel relation and long-term dependence with accurate location information, and divides it into two stages, coordinate information embedding and coordinate attention generation.

(1) Coordinate Information Embedding

Squeezing in the SE module is used for global information embedding. Given the feature tensor input $X = [x_1, x_2, \ldots, x_c,] \in R^{H \times W \times C}$ in the network, the squeezing step for the cth channel can be formulated as follows.

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{1}$$

where z_c is the output associated with the c_{th} channel. H and W are the height and width of the input data feature map. The input X comes directly from a convolution layer with a fixed convolution kernel, and the feature tensor set is obtained by convolution processing.

To enable the attention block to spatially capture long-distance interactions with precise location information, the global pool is decomposed into equations that are converted into one-to-one feature encoding operations. Given an input X, we encode each channel using pooling kernels of sizes $(H, 1)$ or (1, W) along horizontal and vertical coordinates, respectively. Therefore, the output of the c_{th} channel at height h and width w is obtained respectively, and the formula is as follows:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \tag{2}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{3}$$

These two transformations combine features in each of two spatial directions, resulting in a pair of discriminative features with orientation. This is very different from the SE block of the channel attention method, which generates a single feature vector. These two transformations also help the network to locate objects of interest more accurately, which allows the attention module to capture long-term correlation in one space direction and maintain precise position information in another space.

(2) Coordinate Attention Generation

Through the above transformation, we can get a good global perception and encode accurate location information. In order to use the resulting representation, the second transformation is proposed. Given the generated aggregated feature maps, they are connected by Equations (2) and (3) and then transformed using the 1×1 convolutional transform function F_1.

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{4}$$

where $f \in R^{C/r \times (H+W)}$ represents an intermediate feature map encoding spatial information in the horizontal and vertical directions, respectively. Here, r is the reduction ratio used to control the size of the blocks in the SE block. $[\cdot, \cdot]$ represents cascading operations in two spatial dimensions, and δ is a non-linear activation function. Then, we split f into two independent tensors $f^h \in R^{C/r \times H}$ and $f^w \in R^{C/r \times W}$ along the two spatial dimensions of h and w. Using the other two 1×1 convolutions transforms F_h and F_w, and f^h and f^w are transformed into tensors with the same number of input X channels, respectively, as follows:

$$\begin{aligned} g^h &= \sigma\left(F_h\left(f^h\right)\right) \\ g^w &= \sigma(F_w(f^w)) \end{aligned} \tag{5}$$

Here σ is the Sigmoid function, and, after the calculation of Formula (5), the attention weight g^h and g^w of the input feature map in the height direction and in the width direction

will be obtained. Finally, by multiplying and weighting the original feature map, the final feature map with attention weight in the width and height directions will be obtained. The output Y of the coordinate attention block can be written as:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{6}$$

2.4. Resampling Based on Multi-Window Settings

In medical imaging, the principle of MR imaging differs from that of CT. Compared with CT, MR has a better imaging effect on the body's soft tissue and can provide more information. However, there is no corresponding window width and window level setting for MR images of different soft tissues of the body. Therefore, professional radiologists need to adjust each image to a suitable window for lesion labeling when labeling the lesion area, and obtain a better contrast image under this optimal window width and window level. However, the best window position selection for the same body part is usually different under different doctors, machines, sequences, and angle processing. Due to the diversity of window width and window level selection, we believe that the same MR image has different feature information with varying importance in different windows. If we can fuse the information in multiple windows to obtain richer image features, it is beneficial for the deep learning algorithm to analyze image features. From the above analysis, on the one hand, because of the special characteristics of MR images, NPC in MR images cannot identify a relatively fixed optimal window area as CT images do; on the other hand, selecting information within any single window may lead to information loss.

In this paper, we need to perform lesion detection on NPC MR images, and we need to convert the DICOM images to JPG format for neural network training. Traditional medical image processing methods only acquire the image information under a certain window, which leads to a large amount of information loss in the NPC images, so we fuse the image information under multiple windows to obtain a richly layered image for the detector training. In order to improve the data utilization efficiency of the detector for the original image, transmit more image information to the deep learning model, and enable it to obtain richer image features, we adopt the image resampling method based on multi-window setting (our previous MWSR research [30]). The DICOM image metadata is used to obtain information about the preset window (default window width and position), and then the other two windows to the left and right of the preset window are used to obtain a three-channel pseudo-color image, resulting in a more informative and better contrasted nasopharyngeal cancer image.

The image resampling based on multi-window setting is as follows: obtain the preset best window width/level information (ww_0, wl_0) from the MR image metadata of nasopharyngeal cancer in DICOM format. Based on the optimal window (ww_0, wl_0), we set two new window width/level at a certain proportion in the gray level range covered nearby, namely:

$$ww_i = \mu \times ww_0 \; wl_i = \mu \times wl_0, \tag{7}$$

where ww_i and wl_i denote the new window width and window level, respectively, and μ is the weighting factor of the window width and window level.

By observing the image contrast ratio, it was found that the image contrast effect was best at $\mu = 0.5,\ 1.5$, thus two new windows (ww_1, wl_1) and (ww_2, wl_2) were obtained, and the grayscale images under the three windows were combined into RGB images as R, G, and B channels, respectively, to enrich the image information.

As shown in Figure 4, from the range of pixels contained in the MR image (a), the images (b) under the window width/level acquired at μ = 0.5, 1, 1.5 are taken out and displayed according to their grey scale, respectively, as R, G, and B channels to synthesize RGB pseudo-color images (c). Specifically, we convert the NPC MR image (Dicom format) in the data set, and convert the Dicom image into a grayscale image I_0, I_1, I_2 (JPG format) by the above three window width window levels (ww_0, wl_0), (ww_1, wl_1), and (ww_2, wl_2), and then the grayscale images under the three window width window levels are synthesized

into RGB pseudo-color images. The grayscale images under the three window width window levels correspond to the three channels of the RGB image. I_1 corresponds to B channel, I_0 corresponds to G channel, and I_2 corresponds to R channel. It takes about 4 h to process the MR dataset used in this experimental hardware environment, and it is easy and fast to process by computer automation.

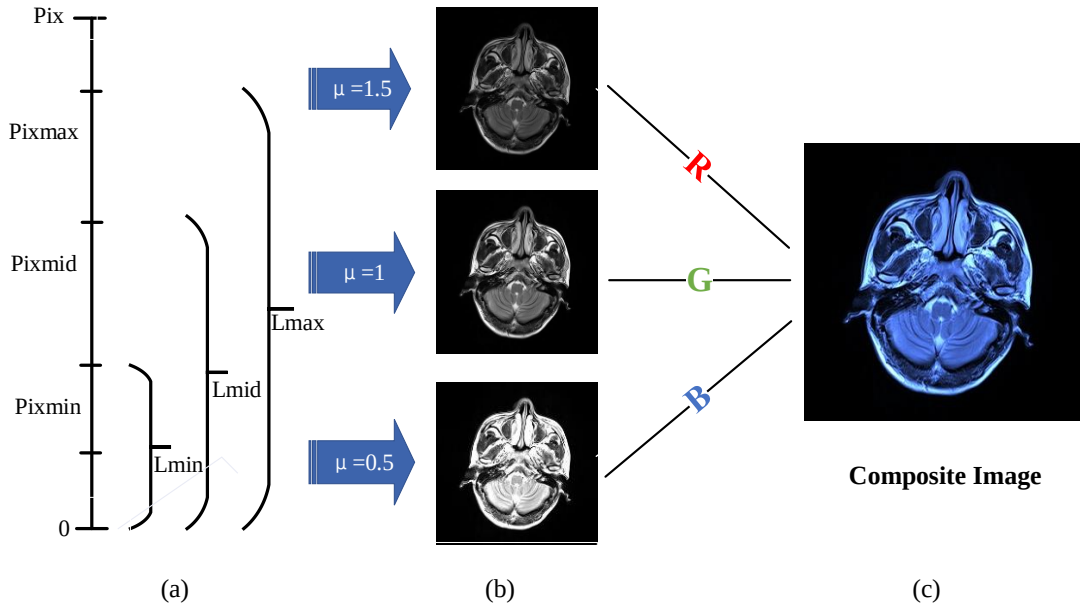


Figure 4. Schematic diagram of MR image resampling based on multi window setting. (**a**) is the pixel range contained in the MR image. (**b**) MR NPC images obtained under three groups of window width and window level. (**c**) is a synthesized RGB pseudo-color image.

2.5. YLCA Network

Considering that the attention mechanism can make the neural network focus on calculating the most important feature vectors, and by embedding position information into the channel attention CA [23], it can not only capture cross-channel information, but also capture direction-aware and position-sensitive information and more accurately locate and identify objects of interest. In the MP-2 module, we replace the 1×1 convolution's CBS module with a CA to build the MP-CA module, as shown in Figure 5.

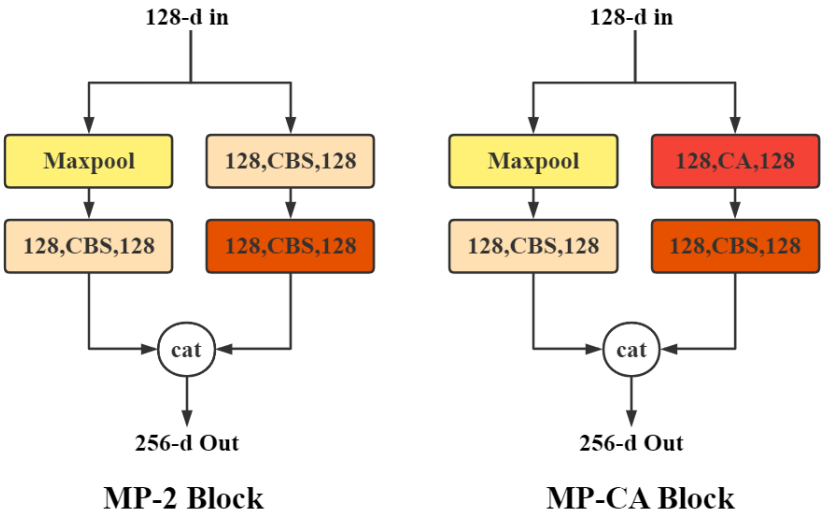


Figure 5. MP-CA Network Module Diagram.

In Figure 5, the CA module uses coordinate attention to process the input features. In YOLOv7, the MP-2 module uses concat to fuse the features extracted from the input data by one pooling, 1 × 1 convolution and one 1 × 1 convolution, and 3 × 1 convolution, respectively. The MP-CA module constructed in this paper replaces the 1 × 1 convolution in the MP-2 block with the CA module, and the input and output remain unchanged. The other structures remain unchanged.

We embedded the fused CA block and MP-CA modules in backbone and head, respectively, in the YOLOv7-based framework, aggregated the primary features extracted from each stage into two independent direction-aware feature maps, encoded them into two attention maps respectively to retain location information, and then applied the two attention maps to the input feature maps to enhance the representation of nasopharyngeal cancer lesion areas to construct a novel network YLCA. A schematic diagram of the YLCA network structure is shown in Figure 6, and this model is used to detect lesion areas of nasopharyngeal cancer.

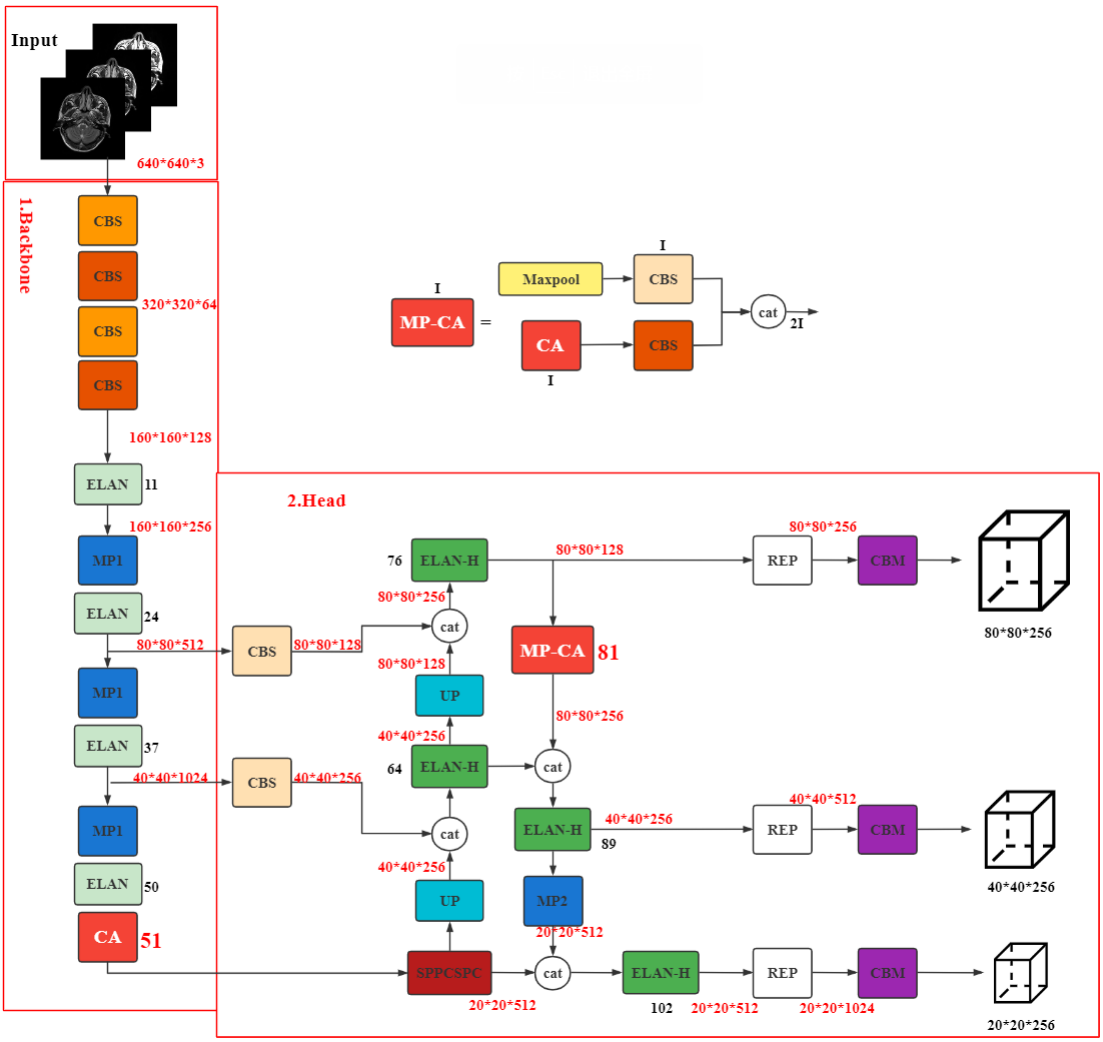


Figure 6. YLCA network structure diagram.

As shown in Figure 6, the YLCA network structure has 107 layers, composed of backbone and head. The first 52 layers are backbone and the last 55 layers are head. In backbone, we add the 51st layer attention mechanism module CA after ELAN, denoted as [−1,1, CoordAtt, [1024]]], where −1 indicates that the upper layer output is the local layer input and the input feature is 1024. In head, we replace the original MP-2 module with the MP-CA module for attention feature fusion at the 81st layer, which is expressed as [−3,1,

CoordAtt, [128]], where −3 indicates that the output of the upper layer 3 is the input of this layer, and the input feature is 128. The schematic diagram of the MA-CA module is also shown in Figure 6, which is used to construct the target object network YLCA.

3. Experimental Settings

3.1. Dataset Description

The experimental data were obtained from the Sun Yat-sen University Cancer Centre. MR images of 800 patients with nasopharyngeal carcinoma were acquired from January 2010 to December 2011. These MR images were T2-weighted (T2WI) axial cross-sectional images with the following imaging parameters: fast spin-echo sequence (FSE), TR = 4000 ms, TE = 99 ms, mean slice thickness of 5 mm, layer spacing of 6 mm, and intraplanar pixel resolution of 0.74 mm × 0.74 mm.

Of the 800 cases, a total of 26,000 MR images were available. Since the nasopharyngeal lesion area accounts for a small proportion of the MR imaging of the head and the clinical presentation is complex and varied, not every image has a lesion area, so only some of the images have a labeled cancerous area. We selected 4694 MR images with lesion areas and corresponding annotated images for the experiment, including 3540 images of male patients and 1154 images of female patients (1596 12-bit DICOM images and 3098 16-bit DICOM images). An expert consensus was formed by four experienced imaging physicians to give the appropriate tumor area annotation. Data enhancement processes, including rotation and horizontal flip, were used. We use the evaluation metrics Precision, Recall, AP (Average Precision), mAP (mean Average Precision), F1-Score, and Confidence to assess the performance of NPC detection.

3.2. Data Conversion

The medical image data used in this experiment is in DICOM format, and the lesion area corresponding to each MR image in the original data is annotated as a PNG format image. In this paper, we use a deep convolutional network for NPC lesion detection, and the annotated area is a rectangular bounding box, so we convert the NPC lesion information outlined by the doctor into rectangular box information and store it in YOLO data annotation format to facilitate the training of the detector. A schematic diagram of the nasopharyngeal cancer lesion labeling process is shown in Figure 7.

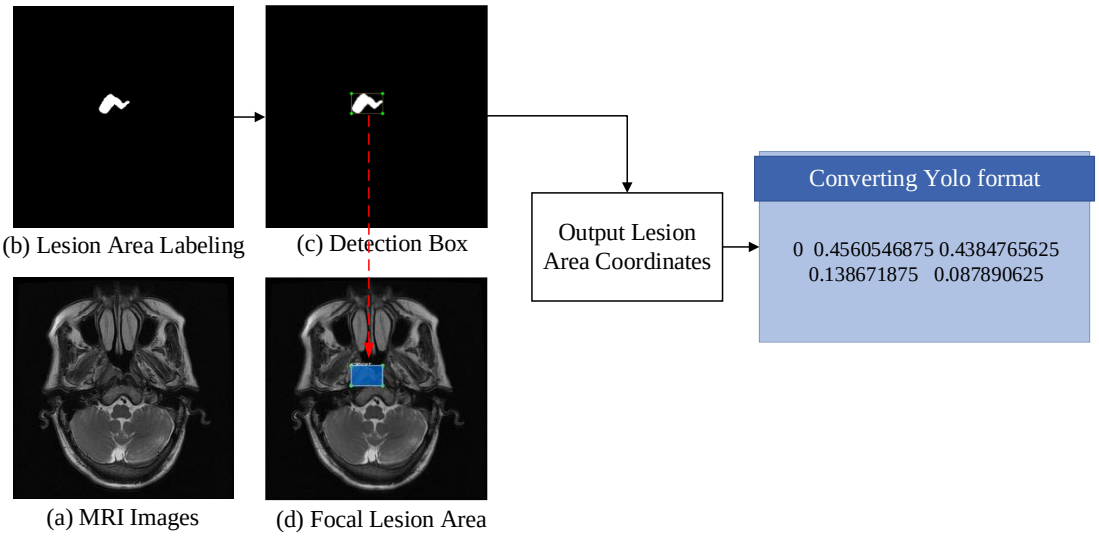


Figure 7. Nasopharyngeal carcinoma lesion labeling processing diagram. (**a**) The original NPC MR image. (**b**) NPC lesion area Labeling. (**c**) The NPC lesion detection Box. (**d**) The real NPC lesion area.

In Figure 7, the image (a) is the original MR medical image. The specific annotation process consists of pixel-level annotation of the nasopharyngeal cancer lesion area (b) by a

pixel traversal algorithm to frame the white lesion area (c) and output the corresponding coordinates and convert the coordinate information into YOLO data format (e), indicating the categories Cancer, X-center, Y-center, w, and h, respectively, then saved as a txt file. The red dashed arrows in the figure point to the effect of converting the pixel-level annotation of the lesion area (white) to (d) a bounding box area covering the real lesion area.

3.3. Other Setting

Data set division: A total of 4694 images were used for detector training in the experiment after data processing. Considering that the detection of NPC lesion area is essentially a single category object detection, and the test set is not easy to be too many, we divided the NPC MR images of each patient into training validation sets and test sets according to 4:1 in 800 patients to ensure the balance of data distribution. Therefore, we used 3755 images as the training validation set, of which 3004 images were used as the training set, 751 images as the validation set, and 939 images as the test set.

Object detector setup: The constructed YLCA network was used as the object detector for this experiment. During the experiment, the input image size was 640 × 640 for both training and testing, the batch size was 32, the initial learning rate was le-2, the loss was calculated using the Stochastic Gradient Descent (SGD) optimizer, the number of iterations of the whole network was 300 epochs. It takes about 2 min to train one iteration (epoch), and each iteration is saved as a model. For the constructed YLCA detector, the performance is optimal around the 180th epoch.

Experimental environment: The algorithm in this paper is built using the deep learning framework Pytorch and the programming language Python. The training and testing of the model is based on the Ubuntu 18.04.6 operating system, with 128G RAM and a high-performance graphics card NVIDIA RTX A6000 GPU (48G).

3.4. Evaluating Metrics

In order to ensure the rationality of the experimental results and the fairness of the comparison test, the algorithm is evaluated by referring to the evaluation indexes widely used in the existing object detection methods, including Precision, Recall, F1, PR curve, and mAP.

(1) Precision rate: the number of samples correctly predicted as true accounts for the proportion of all samples predicted as true, and the accuracy represents the accuracy of the prediction in the positive sample results.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

(2) Recall rate: the number of samples correctly predicted as true accounts for the proportion of all samples that are actually true. In the samples that are actually true, the proportion of predicted positive samples to the total actual positive samples.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

(3) F1-score: F1 is the harmonic mean of precision and recall. It is used to balance the influence of precision and recall, and to evaluate a classifier more comprehensively. The larger F1 indicates the higher quality of the model.

$$F1 = \frac{2 * precision * recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \tag{10}$$

(4) PR curve: according to the value of accuracy and recall rate, the PR curve is drawn to evaluate the model more comprehensively. The larger the area under the PR curve, the higher the average accuracy of the model.

(5) mAP: in this experiment, mAP was used as the main evaluation index. mAP@0.5 (referred: AP_{50}) refers to the value of AP under the condition that IOU (predicted overlap between borders and real borders) is greater than 0.5. mAP@50:5:95 (referred: $AP_{50:95}$) refers to the value of IOU from 0.5 to 0.95, the step size is 0.05, and then the mean value of AP is taken under these IOUs.

4. Results and Discussions

4.1. Evaluation of Multi-Window Resampling

To verify the effectiveness of the resampling method based on the multi-window setting (i.e., multi-window width window setting), we performed object detection experiments on a single-window nasopharyngeal cancer image set (gray) and a multi-window resampled nasopharyngeal cancer image set (MWSR [30]), respectively. The results are shown in Table 1, where P^{test} is the Precision of the test set, and R^{test} is the Recall of the test set. We use the model trained under the YOLOv7 detector using the gray image set as the baseline model.

Table 1. Comparison of model performance on single window and multi window images.

Sample Selection	Model	P^{test}	R^{test}	AP^{val}_{50}	$AP^{val}_{50:95}$	AP^{test}_{50}	$AP^{test}_{50:95}$
Gray	YOLOv7	0.771	0.72	77.0%	36.0%	77.8%	36.7%
MWSR	YOLOv7	0.812	0.714	78.3%	36.3%	78.9%	36.8%
Gray	YLCA (Ours)	0.802	0.719	78.0%	**37.0%**	78.3%	36.5%
MWSR	YLCA (Ours)	0.839	0.711	**79.0%**	36.4%	**80.1%**	**37.6%**

From Table 1, it can be seen that when using the YOLOv7 object detector for experiments, compared to the single-window NPC image set (gray), the multi-window setting-based resampled RGB image (MWSR) object detection performance is better, and the evaluation metrics of AP^{val}_{50}, $AP^{val}_{50:95}$, AP^{test}_{50}, and $AP^{test}_{50:95}$ are improved by 1.3%, 0.3%, 1.1%, and 0.1%, respectively. In experiments based on our YLCA object detector, the MWSR improved the evaluation metrics of AP^{val}_{50}, AP^{test}_{50}, and $AP^{test}_{50:95}$ by 1.0%, 1.8%, and 1.1% respectively. In this experiment, the NPC MR image set with multi-window setting is tested under different models. It is verified that the resampling processing based on multi-window setting for NPC MR image can improve the data utilization rate of the original image, and the detection performance is improved under different models, which is suitable for improving the lesion detection performance of NPC tumor.

4.2. Ablation Study

The algorithm in this paper mainly includes two parts: image resampling and constructing YLCA network. The ablation study can be divided into five parts: MWSR, gray, YOLOv7, CA and MP-CA. They are respectively using training data MWSR or gray, using object detector original YOLOv7 model, and adding fusion attention mechanism network CA and MP-CA modules, respectively.

As shown in Table 2, with the addition of each module, AP_{50} and $AP_{50:95}$ gradually increase, indicating that each module can improve the detection ability of the network. Compared with (1) and (2) in Table 2, it can be seen that when using the YOLOv7 object detector for experiments, the RGB map (MWSR) object detection performance based on multi-window image resampling is better. Compared with the single-window NPC image set (gray), the AP is improved by 1.1% and 0.1%, respectively. Compared with (2) and (3) in Table 2, the AP is improved by 0.6% and 0.4%, respectively, after adding the attention feature module (CA) compared with the simple YOLOv7 network. Compared with (2) and (4) in Table 2, the AP is improved by 0.3% after adding the attention feature fusion module (MP-CA). Compared with (2) and (6), the attention mechanism and attention feature fusion module are combined and applied to the YOLOv7 network, and the AP is significantly improved, reaching 1.2% and 0.8%, respectively.

Table 2. Quantitative test results for each module of the MWSR-YLCA algorithm.

Number	MWSR	Gray	YOLOv7	CA	MP-CA	AP_{50}^{test}	$AP_{50:95}^{test}$
(1)		√	√			77.8%	36.7%
(2)	√		√			78.9%	36.8%
(3)	√		√	√		79.5%	37.2%
(4)	√		√		√	79.2%	37.1%
(5)	√		√	√	√	**80.1%**	**37.6%**

4.3. Algorithm Comparison and Analysis

In order to verify the effectiveness of the proposed algorithm, it is compared with some mainstream object detection algorithms. Due to the particularity of NPC data sets, the data required by different models are different. Here, it is mainly compared with the newer YOLO series detection models, including: YOLOv5 [31], YOLOR [32], YOLOX [33], YOLOv7 [22], and RetinaNet [21]. In order to compare different experiments fairly, we used multi-window resampling NPC image set for training, and used the optimal model provided by the original paper for testing.

We compare the proposed method with the state-of-the-art object detectors. From Table 3, we can see that YLCA performs best in the NPC MR dataset compared with the five newer object detection algorithms (using their respective optimal models). Our attention fusion network enhances the feature representation of NPC lesion area and improves the object detection performance. Our network achieves the highest detection performance on F1, AP_{50}^{val}, AP_{50}^{test}, and $AP_{50:95}^{test}$, and the highest $AP_{50:95}^{val}$ appears on YOLOv7-tiny. Compared with the current excellent YOLOv7, the AP of our method is 0.3%, 0.4%, 1.2%, and 0.8% higher, respectively. In addition, the AP_{50}^{test} and $AP_{50:95}^{test}$ of our model in the test set are 0.9% and 0.8% higher than the best models YOLOv7-tiny and YOLOv7, respectively. The model training time of YLCA is about 8.2 h, and the FPS value is 77.52. The training and reasoning speed is close to that of YOLOv7, but the AP is increased by 1.2%, which suggests that our method improves performance while maintaining training rate. As can be seen from Table 3, our YLCA achieves the highest indicator F1 of 0.77 and a higher indicator P^{test} of 0.839, which has a substantial improvement compared to other methods and has a better PR curve. This shows that the prediction image obtained in the large-scale dataset is closer to the truth value and has higher average accuracy. The P-R and F1 curves of the YLCA model are shown in Figure 8.

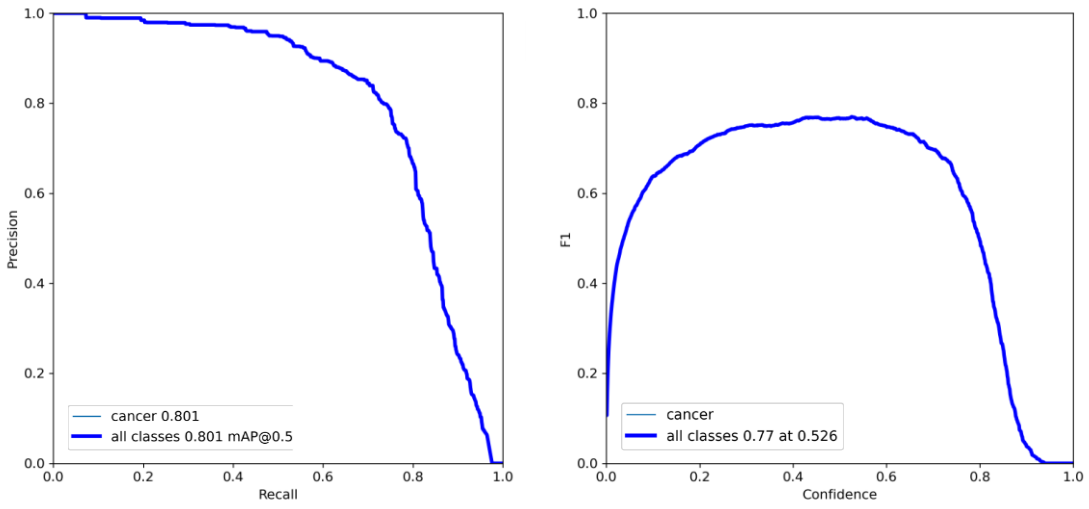


Figure 8. P-R and F1 curves of YLCA model.

Table 3. Comparison of lesion detection results based on each model of NPC MR dataset.

Model	Size	P^{test}	R^{test}	F1	AP^{val}_{50}	$AP^{val}_{50:95}$	AP^{test}_{50}	$AP^{test}_{50:95}$	Year
YOLOv5-S [31]	640	0.809	0.717	0.76	76.0%	34.0%	77.7%	35.0%	2021
YOLOv5-X [31]	640	0.767	0.723	0.74	76.0%	34.2%	77.0%	34.5%	2021
YOLOv5-L [31]	640	0.781	0.681	0.73	76.1%	34.1%	74.0%	33.1%	2021
YOLOv5-M [31]	640	0.757	0.754	0.76	75.6%	35.7%	77.6%	35.8%	2021
YOLOR-CSP [32]	640	0.786	0.708	0.75	76.0%	35.7%	77.0%	35.5%	2021
YOLOR-CSP-X [32]	640	0.8	0.715	0.75	77.4%	35.3%	78.0%	36.8%	2021
YOLOR-P6 [32]	640	0.761	0.668	0.71	69.3%	31.0%	71.5%	30.8%	2021
YOLOR-D6 [32]	640	0.771	0.672	0.72	71.4%	31.7%	72.7%	33.0%	2021
YOLOX-S [33]	640	0.855	0.657	0.74	76.9%	35.7%	78.2%	36.4%	2021
YOLOX-X [33]	640	0.817	0.705	0.76	75.1%	33.5%	75.5%	33.2%	2021
YOLOX-L [33]	640	0.838	0.683	0.75	76.7%	34.3%	76.4%	34.1%	2021
YOLOX-M [33]	640	0.842	0.666	0.74	76.5%	34.9%	76.3%	34.7%	2021
YOLOv7 [22]	640	0.812	0.714	0.76	78.3%	36.0%	78.9%	36.8%	2022
YOLOv7-X [22]	640	0.812	0.707	0.76	77.0%	35.9%	78.4%	36.2%	2022
YOLOv7-tiny [22]	640	0.804	0.737	0.77	77.8%	**37.2%**	79.2%	36.5%	2022
YOLOv7-W6 [22]	640	0.814	0.678	0.74	76.9%	35.3%	77.3%	35.5%	2022
YOLOv7-E6 [22]	640	0.807	0.686	0.74	75.6%	35.0%	76.9%	35.8%	2022
YOLOv7-D6 [22]	640	0.785	0.706	0.74	76.4%	34.8%	77.8%	36.2%	2022
YOLOv7-E6E [22]	640	0.803	0.704	0.75	75.8%	34.8%	77.1%	35.8%	2022
RetinaNet [21]	640	-	0.913	-	-	-	72.9%	-	2018
YLCA (Ours)	640	0.839	0.711	**0.77**	**79.0%**	36.4%	**80.1%**	**37.6%**	2022

In terms of statistical tests, we performed the Shapiro–Wilk test on all the test data, and the test results showed that our data were normal. Therefore, we use t-test to determine whether there is a significant difference between the mean values of the two variables, to let us know whether they belong to the same distribution. We conducted five experiments under MWSR-CAYL and baseline detectors, and performed *t*-tests under AP^{test}_{50} and $AP^{test}_{50:95}$, respectively. The null hypothesis is that the mean values of the two groups of data are the same. We set a high standard of Alpha = 0.01 for testing, and the *p*-values under the two indicators are 0.0021 and 0.0024, respectively. The experimental results show that the *p*-value is less than the Alpha value under both detection indicators. We will reject the null hypothesis and show that our data is statistically significant.

To ensure that the experimental results are reasonable, the algorithm is evaluated with reference to the confusion matrix that is widely used in existing object detection methods. The confusion matrix for model evaluation is the most basic, intuitive, and computationally simple method to measure the accuracy of the sub-types of models. Where the horizontal coordinates represent the true labels and the vertical coordinates represent the predicted results of the model, the confusion matrix of the YLCA model on MR images of nasopharyngeal carcinoma is shown in Figure 9.

Object detection determines whether the test results are correct. The most commonly used way is to calculate the detection box and the real box IOU, and then, according to the IOU, determine whether the two boxes match. As shown in Figure 9, the full predictions of the test set of NPC MR are evaluated, and the four squares are True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN). The confusion matrix of object detection differs from classification in that it focuses primarily on the detection result of the detection box rather than the entire image. Among them, TP indicates that the model's detection result of NPC MR data is cancer, and its true label is also the ratio of cancer. The ratio of correctly detecting cancer is 0.83. FP indicates that the detection box has positioning or classification errors, the true label is background, and the ratio of model detection results to cancer is 1. FN indicates that the model is missed, and the ratio of the model not

detecting true label cancer is 0.17. Since the detector does not detect a background area, TN is represented as 0.

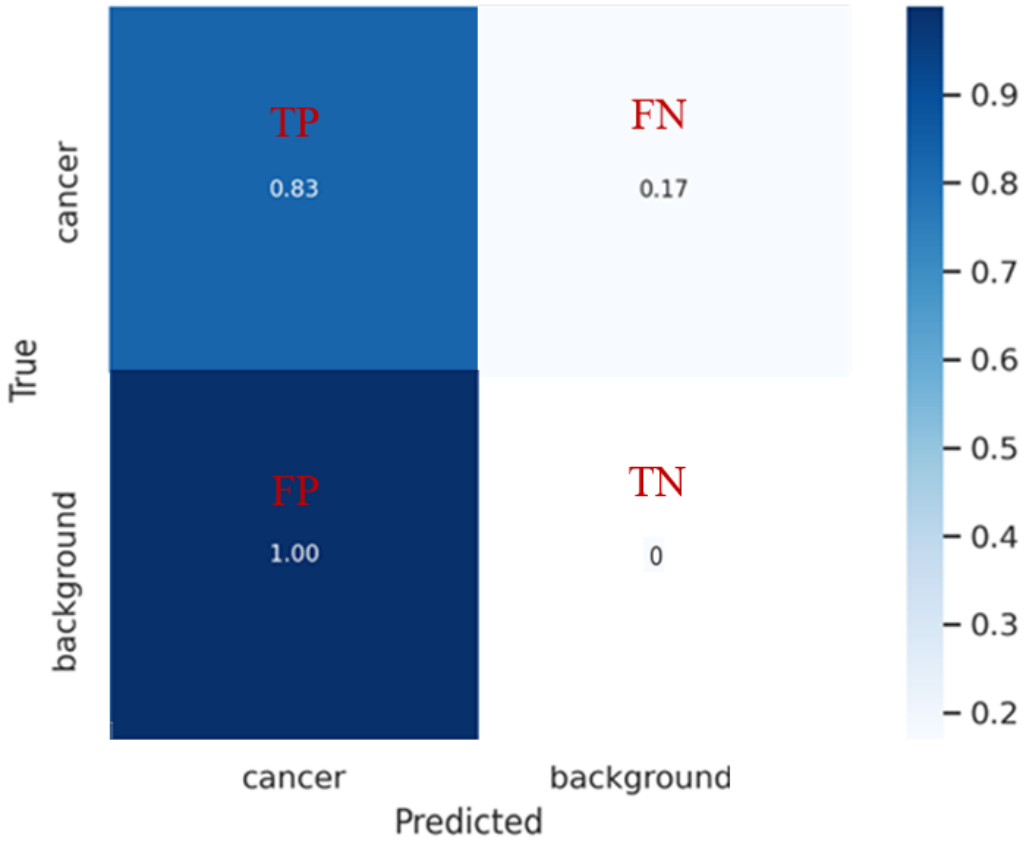


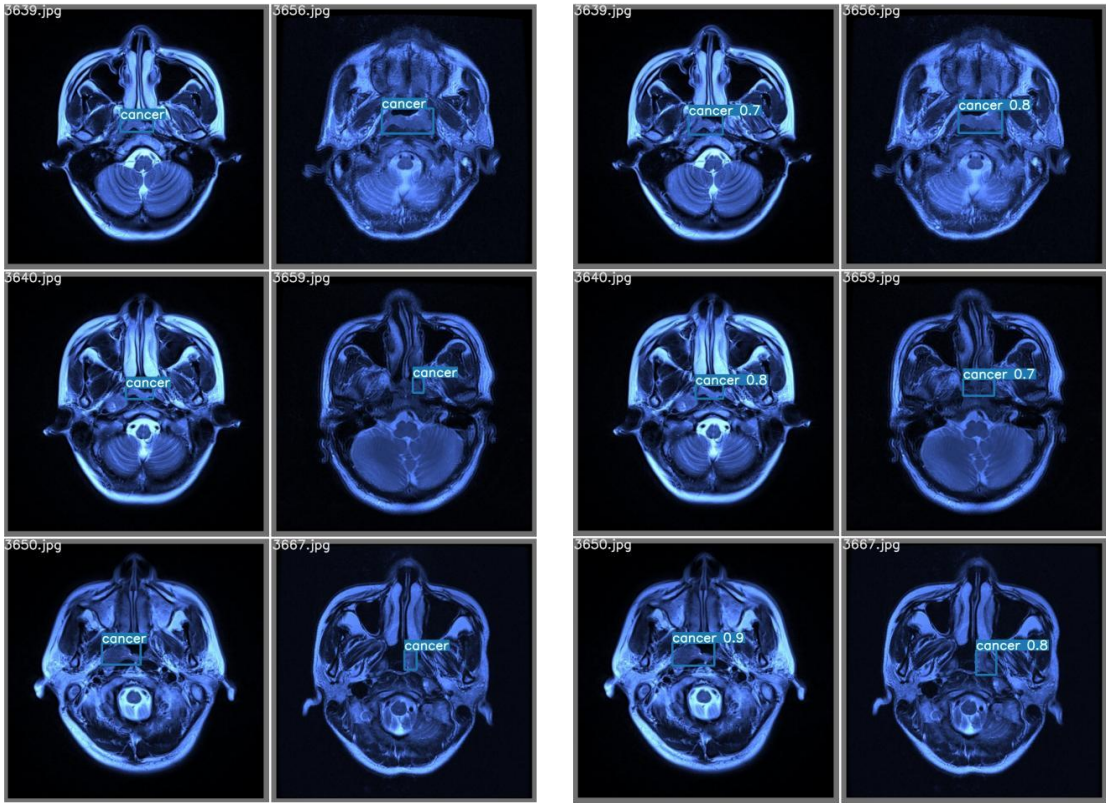
Figure 9. Confusion matrix of YLCA model on MR images of nasopharyngeal carcinoma.

Figure 10 shows the performance of YLCA network model in NPC MR images. The 6 sample images (Figure 10a) were selected randomly from the NPC MR test set. It is an NPC MR image of different patients at different angles. In Figure 10, (a) was a NPC image with labeled cancer area, which represented the real lesion area and category, and (b) was the prediction result of the corresponding image predicted by the detector, which represented the predicted lesion area, category, and confidence. From the results, by comparing Figure 10a,b our detector was relatively accurate in detecting the lesion area. The detector's prediction of NPC images is close to the true lesion area and has a high classification confidence. The algorithm in this paper can use image features more efficiently to detect smaller NPC lesion area. The overall performance shows that the prediction map of the object detection algorithm embedded in the attention mechanism fusion network can better approach the truth map.

4.4. Discussions

In this experiment, the performance of the NPC detection process is steadily improved by multi-window resampling of the NPC data set. Firstly, our resampling method based on multi-window setting improves the detection mAP (AP_{50}^{test}) of YOLOv7 and YLCA network by 1.1% and 1.8%, respectively, on the test set, which indicates that our method reduces the information loss of NPC MR images and improves the data utilization rate, which is meaningful for medical tumor data that is difficult to obtain in large quantities. Then, we perform object detection under NPC MR images, introduce the CA attention mechanism to construct a YLCA network for NPC lesion detection, and in the case of the same NPC data processed by MWSR for comparative experiments, the mAP of our YLCA network is 1.2% higher than that of YOLOv7, which indicates that our method can make the detection

attention of the model focus on NPC lesions and surrounding tissues. In general, our MWSR-YLCA method performs well on the test set, and the performance is 2.3% higher than the baseline model (AP_{50}^{test}) in the first row and the fourth row of Table 1. This is an improvement for computer-aided detection algorithm of NPC, which shows that our algorithm is very effective. Compared with the latest YOLO detector, our algorithm obtains the highest mAP of 80.1%. Due to the different performance of different models in the data set, our algorithm is more accurate in the latest NPC detection method.



（a）True lesion area （b）Predicted lesion area

Figure 10. The manifestation of YLCA on MR images of nasopharyngeal carcinoma. (**a**) NPC images of the labeled cancer area. (**b**) The prediction result of the corresponding image predicted for the detector.

In addition, from the practical application, we can directly use the annotated MR data of clinical imaging experts for NPC detection, and the data processing method is simple and efficient, and the time and calculation cost are more convenient and faster than the existing NPC detection methods that are mostly complex modalities. On the dataset, we use NPC images of 800 patients (4694 annotated slices), which is also abundant in the existing NPC research and has certain generalization ability. In summary, our algorithm has superior performance and can achieve high-performance NPC detection.

5. Conclusions

In this paper, we propose a computer-aided detection method (MWSR-YLCA) for NPC lesion detection in MR images. Specifically, we design two modules, the multi-window settings resampling (MWSR) module and an improved YOLOv7 with embedded a coordinate attention mechanism (YLCA) module, to detect NPC lesions more accurately. Firstly, the NPC MR image is resampled based on the MWSR module. The comparison experiments show that this method can fuse the feature information of medical MR image with a deep learning network more effectively and enhance the information utilization.

On this basis, a detection network YLCA with embedded fusion attention mechanism is constructed to detect NPC lesions. Qualitative and quantitative comparative analysis and ablation experiments are carried out with other algorithms. The results show that the CA module can effectively extract lesion features, and YLCA network has better performance for lesion detection. In summary, the MWSR-YLCA is capable of performing highly accurate detection of NPC lesions and has good performance and important applications.

Author Contributions: Conceptualization, H.W., X.Z., G.H., H.L., Y.K. and J.L.; methodology, H.W., X.Z. and G.H.; validation, H.W., X.Z. and G.H.; formal analysis, H.W.; resources, G.H. and H.L.; visualization, H.W. and X.Z.; supervision, G.H. and H.L.; project administration, G.H. and H.L.; funding acquisition, G.H.; writing—original draft preparation, H.W. and X.Z.; writing—review and editing, H.W., X.Z., G.H., H.L., Y.K. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (61901533), Shenzhen Fundamental Research Program, China (JCYJ20190807154601663), and High-level Talents Research Project of NCWU (202101002).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki. Ethical review and approval were waived for this study, due to the retrospective nature of the survey.

Informed Consent Statement: Patient consent was waived due to the retrospective design of this study.

Data Availability Statement: The MRI image data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wei, W.I.; Sham, J.S.T. Nasopharyngeal carcinoma. *Lancet* **2005**, *365*, 2041–2054. [CrossRef] [PubMed]
2. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
3. Li, H.; Cao, D.; Li, S.; Chen, B.; Zhang, Y.; Zhu, Y.; Luo, C.; Lin, W.; Huang, W.; Ruan, G.; et al. Synergistic Association of Hepatitis B Surface Antigen and Plasma Epstein-Barr Virus DNA Load on Distant Metastasis in Patients with Nasopharyngeal Carcinoma. *JAMA Netw. Open* **2023**, *6*, e2253832. [CrossRef]
4. Abdulhay, E.; Mohammed, M.A.; Ibrahim, D.A.; Arunkumar, N.; Venkatraman, V. Computer aided solution for automatic segmenting and measurements of blood leucocytes using static microscope images. *J. Med. Syst.* **2018**, *42*, 58. [CrossRef]
5. Huang, W.; Chan, K.L.; Zhou, J. Region-based nasopharyngeal carcinoma lesion segmentation from MRI using clustering-and classification-based methods with learning. *J. Digit. Imaging* **2013**, *26*, 472–482. [CrossRef] [PubMed]
6. Mohammed, M.A.; Abd Ghani, M.K.; Arunkumar, N.A.; Mostafa, S.A.; Abdullah, M.K.; Burhanuddin, M.A. Trainable model for segmenting and identifying Nasopharyngeal carcinoma. *Comput. Electr. Eng.* **2018**, *71*, 372–387. [CrossRef]
7. Huang, B.; Chen, Z.; Wu, P.M.; Ye, Y.; Feng, S.T.; Wong, C.Y.; Zheng, L.; Liu, Y.; Wang, T.; Li, Q.; et al. Fully automated delineation of gross tumor volume for head and neck cancer on PET-CT using deep learning: A dual-center study. *Contrast Media Mol. Imaging* **2018**, *2018*, 8923028. [CrossRef]
8. Chen, H.; Qi, Y.; Yin, Y.; Li, T.; Liu, X.; Li, X.; Gong, G.; Wang, L. MMFNet: A multi-modality MRI fusion network for segmentation of nasopharyngeal carcinoma. *Neurocomputing* **2020**, *394*, 27–40. [CrossRef]
9. Wang, S.; Zhu, Y.; Lee, S.; Elton, D.C.; Shen, T.C.; Tang, Y.; Peng, Y.; Lu, Z.; Summers, R.M. Global-Local attention network with multi-task uncertainty loss for abnormal lymph node detection in MR images. *Med. Image Anal.* **2022**, *77*, 102345. [CrossRef]
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
11. Huang, K.W.; Zhao, Z.Y.; Gong, Q.; Zha, J.; Chen, L.; Yang, R. Nasopharyngeal carcinoma segmentation via HMRF-EM with maximum entropy. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milano, Italy, 25–29 August 2015.
12. Li, S.; Xiao, J.; He, L.; Peng, X.; Yuan, X. The tumor target segmentation of nasopharyngeal cancer in CT images based on deep learning methods. *Technol. Cancer Res. Treat.* **2019**, *18*, 1533033819884561. [CrossRef]
13. Alom, M.Z.; Aspiras, T.; Taha, T.M.; Asari, V.K.; Bowen, T.J.; Billiter, D.; Arkell, S. Advanced deep convolutional neural network approaches for digital pathology image analysis: A comprehensive evaluation with different use cases. *arXiv* **2019**, arXiv:1904.09075.

14. Liu, L.; Chen, S.; Zhang, F.; Wu, F.X.; Pan, Y.; Wang, J. Deep convolutional neural network for automatically segmenting acute ischemic stroke lesion in multi-modality MRI. *Neural Comput. Appl.* **2020**, *32*, 6545–6558. [CrossRef]
15. Zhang, M.; Young, G.S.; Chen, H.; Li, J.; Qin, L.; McFaline-Figueroa, J.R.; Reardon, D.A.; Cao, X.; Wu, X.; Xu, X. Deep-learning detection of cancer metastases to the brain on MRI. *J. Magn. Reson. Imaging* **2020**, *52*, 1227–1236. [CrossRef] [PubMed]
16. Elakkiya, R.; Teja, K.S.; Jegatha Deborah, L.; Bisogni, C.; Medaglia, C. Imaging based cervical cancer diagnostics using small object detection-generative adversarial networks. *Multimed. Tools Appl.* **2022**, *81*, 191–207. [CrossRef]
17. Salman, M.E.; Çakar, G.Ç.; Azimjonov, J.; Kösem, M.; Cedïmoğlu, İ.H. Automated prostate cancer grading and diagnosis system using deep learning-based Yolo object detection algorithm. *Expert Syst. Appl.* **2022**, *201*, 117148. [CrossRef]
18. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
21. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 318–327. [CrossRef]
22. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
23. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
24. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
25. Janiesch, C.; Zschech, P.; Heinrich, K. Machine learning and deep learning. *Electron. Mark.* **2021**, *31*, 685–695. [CrossRef]
26. Jie, H.; Li, S.; Gang, S.; Albanie, S. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *42*, 2011–2023.
27. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
28. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
29. Liu, N.; Zhang, N.; Han, J. Learning Selective Self-Mutual Attention for RGB-D Saliency Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
30. Han, G.; Liu, X.; Zhang, H.; Zheng, G.; Soomro, N.Q.; Wang, M.; Liu, W. Hybrid resampling and multi-feature fusion for automatic recognition of cavity imaging sign in lung CT. *Future Gener. Comput. Syst.* **2019**, *99*, 558–570. [CrossRef]
31. Jocher, G.; Stoken, A.; Borovec, J.; Chaurasia, A.; Changyu, L.; Laughing, A.; Hogan, A.; Hajek, J.; Diaconu, L.; Kwon, Y.; et al. ultralytics/yolov5: v5. 0-YOLOv5-P6 1280 models AWS Supervise. ly and YouTube integrations. *Zenodo* **2021**. [CrossRef]
32. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. You only learn one representation: Unified network for multiple tasks. *arXiv* **2021**, arXiv:2105.04206.
33. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.