future internet

Article

Single-Shot Global and Local Context Refinement Neural Network for Head Detection

Jingyuan Hu * and Zhouwang Yang

School of Mathematical Sciences, University of Science and Technology of China, No. 96, JinZhai Road Baohe District, Hefei 230026, China

* Correspondence: hjyq@mail.ustc.edu.cn

Abstract: Head detection is a fundamental task, and it plays an important role in many head-related problems. The difficulty in creating the local and global context in the face of significant lighting, orientation, and occlusion uncertainty, among other factors, still makes this task a remarkable challenge. To tackle these problems, this paper proposes an effective detector, the Context Refinement Network (CRN), that captures not only the refined global context but also the enhanced local context. We use simplified non-local (SNL) blocks at hierarchical features, which can successfully establish long-range dependencies between heads to improve the capability of building the global context. We suggest a multi-scale dilated convolutional module for the local context surrounding heads that extracts local context from various head characteristics. In comparison to other models, our method outperforms them on the Brainwash and the HollywoodHeads datasets.

Keywords: head detection; one-stage detector; convolutional neural network; global context; local context

check for updates

Citation: Hu, J.; Yang, Z. Single-Shot Global and Local Context Refinement Neural Network for Head Detection. *Future Internet* **2022**, *14*, 384. https://doi.org/10.3390/fi14120384

Academic Editor: Young Im Cho

Received: 23 November 2022
Accepted: 14 December 2022
Published: 19 December 2022

1. Introduction

Numerous visual investigations, including object tracking [1,2], person identification [3], face recognition [4,5], and crowd understanding [6], depend on head detection. Locating every person and counting them is crucial in many security and emergency-management scenarios in restaurants, conference centers, and subway stations. The occlusion of the target body part, however, limits the ability to detect people. Face detection also seems to be helpless when the person is facing away from the camera. These factors make head detection more suitable for locating and counting people than person detection and face detection. However, the wide variations in body attitudes, orientations, occlusions, and lighting conditions make this task a remarkable challenge even today.

Without taking into account the characteristics of head detection, many current methods view head detection as a specific case of generic object detection. CNN-based object detectors, represented by Faster-RCNN [7], SSD [8], YOLO [9], and their variants, have significantly improved performance of generic object detection. Generally, these methods use learned models or predetermined anchors to predict proposals first, and then they extract the features of the chosen proposals for additional classification and regression. However, these methods deal with candidate boxes separately. Furthermore, they only extract features from the appearance near the objects' region of interest, ignoring a lot of the context information, such as object–scene relationships and object–object similarity in the same lighting condition. Consequently, it is challenging to apply these models directly to head detection, especially in complex scenes.

To address these problems, Vu et al. [10] introduce a context-aware method for detecting heads. Initially, this model predicts candidate boxes. The context-aware model then explicitly creates relationships between people using a pairwise model and establishes object–scene relations by learning a global CNN model. Their context-aware model achieves impressive performance compared with traditional models, for instance, DPM [11].

The context-aware model, however, has a number of drawbacks, including the following: (1) capturing global context by repeated convolution layers is difficult to optimize [12]; (2) as the quantity of targets rises, the computational complexity of the model is amplified exponentially; and (3) local context information surrounding objects is ignored.

After that, Li et al. [13] present HeadNet, which achieves much better performance. In order to obtain local context information, HeadNet estimates similarity around the object and combines the results from layers with various receptive fields. However, there are three issues with HeadNet: (1) the local context estimate is computationally sophisticated; (2) the merge strategy is based on prior statistical analysis; and (3) long-range dependency is underutilized.

In this paper, to tackle the problems mentioned above, we propose a novel one-stage head detector called Context Refinement Network (CRN), which refines both the local and global context without any prior knowledge and identifies heads via their center points in the images. People's heads perform similar features in the same environment, which is the motivation behind our method, and contextual information can increase classification accuracy. We introduce two context-refinement modules called the global context refinement module (GRM) and the local context refinement module (LRM), respectively. GRM captures the global context by explicitly establishing long-distance dependencies between pixels of the feature map. In addition, the global context can be further refined by fusing feature maps from different stages. Then, LRM is designed to obtain the local context around objects by enlarging the receptive field through multi-scale dilated convolution while maintaining a low computational complexity.

Our model outperformed other models with a 93.89% AP on the Brainwash dataset and an 83.4% AP on the HollywoodHeads dataset.

The following is a summary of our contributions:

- We designed a one-stage head detector, namely, Context Refinement Network (CRN), which is end-to-end trainable without any artificial priors.
- We present the global context refinement module (GRM), which creates long-range dependencies between heads to improve the global context.
- We propose a local context refinement module (LRM) with a multi-scale architecture to refine the local context.
- Our proposed head detector significantly improve the performance on the Brainwash and the HollywoodHeads datasets.

2. Related Work

2.1. Generic Object Detection

Deep learning has made immeasurable progress in the fields of generic object detection in recent years. The one-stage detector and the two-stage detector are the two series that are typically used to classify the currently proposed CNN-based object detectors. The former detectors, represented by YOLO [9], SSD [8], RefineDet [14], RefineDet++ [15], and RetinaNet [16], jointly predict classes and locations of objects without generating candidate boxes, which generally are end-to-end trainable. CornetNet [17] and CenterNet [18] are two recently proposed one-stage detectors, which identify objects via points in the images. Two-stage detectors, including Faster R-CNN [7], Mask R-CNN [19], and Cascade R-CNN [20], present object classification and bounding box regression on the selected proposals after using a region proposal network (RPN) to roughly estimate object proposals. However, because each candidate box is treated separately, these methods can hardly be used to directly solve specific problems [21]. Our proposed method is related to one-stage approaches. However, our method can fuse contextual information around objects via dilated convolution, which is significant important to objects with intra-individual stability. Furthermore, we employ multiple dilated convolution layers in parallel to aggregate the multi-scale context.

2.2. Head Detection

As a subtask of object detection, studies of head detection have benefit from many approaches of generic object detection. Wang et al. [22] present a variant of SSD by concatenating features from different layers. Vu et al. [10] propose a context-aware model via building relations among pairs of objects. Subsequently, Li et al. [13] introduce a fuse strategy, which combines information around object proposals, based on Faster R-CNN [19]. Li et al. [13] find that dilated convolution can effectively extract relationship between the head and shoulders. Our work is related to these approaches in terms of fusing local context. The difference in our method is that we not only model short-range dependencies around objects but also establish long-range dependencies between objects for capturing the global context.

2.3. Pairwise Long-Range Dependency Modeling

Long-range dependency modeling has been extensively studied in many fields, such as video segmentation [23] and image segmentation [24]. The self-attention module is one of the first to model pairwise long-range relations. Furthermore, [25] made an effort to make use of this module to establish long-term relationships between words in machine translation. Furthermore, self-attention mechanisms have been successfully carried over into various other fields, such as generative modeling [26], graph embedding [27], and visual recognition [28,29]. Subsequently, Non-local Networks [30] extend self-attention mechanisms to establish the relations between any two pixels by computing an affinity matrix. However, Cao et al. [12] point out that the Non-local method is computationally challenging and difficult to integrate. Simplified Non-local (SNL) and Global Context (GC) blocks, according to [12], are more efficient and perform better. As a result of these studies, SNL is integrated into our framework to effectively acquire long-range dependency. Furthermore, in order to obtain more refined relationships between objects, we fuse the output feature maps of SNL blocks from various layers.

3. Our Approach

Our network architecture and work flow are firstly described in this section. The use of a multi-level feature map to capture a more precise global context is then demonstrated. Last but not least, we present the local context refinement module, which aids in obtaining an aggregated multi-scale context to improve the overall performance.

3.1. One-Stage Object Detector

Refer to the overall network architecture shown in Figure 1. In our neural network, similar to CenterNet [18], we regard a head as a center point, that is, the center of the bounding box. For objective comparison with most existing methods, our framework adopts ResNet50 [31] as the backbone. This model is trained on the ImageNet dataset. As in HeadNet [13], SANM [32], and FPN [33], all layers after conv5 are removed. We denote the activated feature generated by each stage as $\{S_2, S_3, S_4, S_5\}$. In relation to the input image, $\{S_2, S_3, S_4, S_5\}$ have strides of $\{4, 8, 16, 32\}$, and their channel numbers are $\{256, 512, 1024, \text{and } 2048\}$, respectively. We design a top-down pyramidal structure with GRM, which will be discussed below, to restore resolution from higher stage levels in order to capture high-quality relations between objects. The dimension of final output feature map from a series of GRMs is $64 \times 128 \times 128$. After that, a context-aware module is exploited to fuse the local context information around objects without changing the dimension of the feature map. Finally, three parallel branches are used to estimate the head heatmap, bounding box sizes, and head center offsets, respectively. Each branch is implemented by applying a 1×1 convolution to generate the final targets.

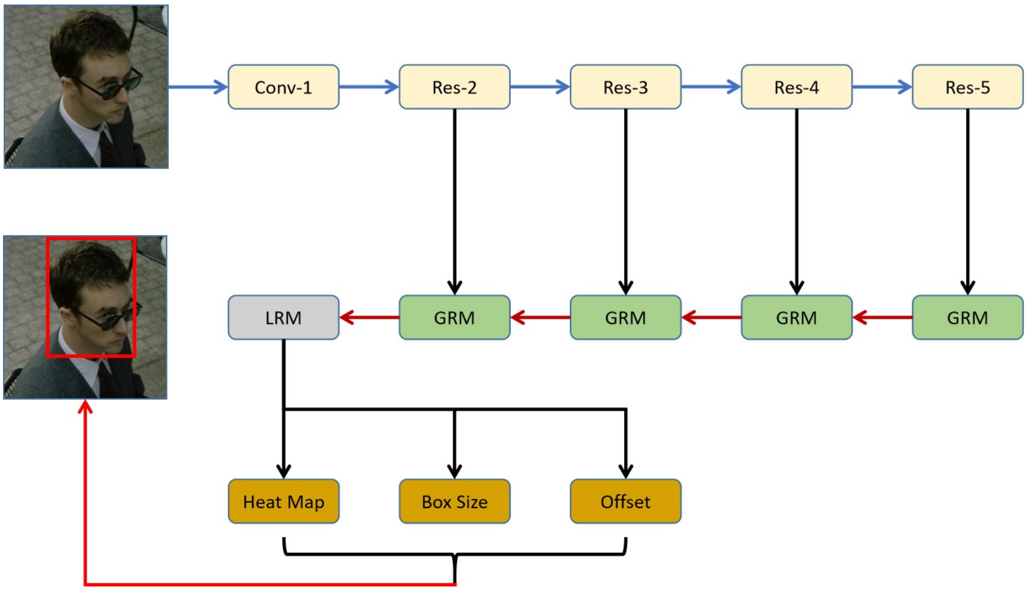

Figure 1. An overview of our CRN architecture: the yellow part is the resnet50.

3.1.1. Heatmap Branch

The heatmap aims to estimate the position of the head center. In particular, for our problem, the shape of the heatmap is $1 \times H \times W$. The value of a location in the heatmap is anticipated to be 1 if it corresponds to the ground-truth head center and 0 for background pixels. As the distance between the heatmap location and the head center increases, the response within the radius exponentially decays, and the background outside the radius is 0. Following [34], we set the radius of each head center point to a constant 2.

For each ground truth box $\mathbf{b}^k = (x_1^k, y_1^k, x_2^k, y_2^k)$, the center of it is $\mathbf{c}^k = (c_x^k, c_y^k)$, where $c_x^k = \frac{x_1^k + x_2^k}{2}$ and $c_y^k = \frac{y_1^k + y_2^k}{2}$. Then its location on the headmap can be computed by dividing the output stride; here, the output stride is set to 4. Therefore, $\hat{\mathbf{c}}^k = (\hat{c}_x^k, \hat{c}_y^k) = (\lfloor \frac{c_x^k}{4} \rfloor, \lfloor \frac{c_y^k}{4} \rfloor)$. After that, the heatmap target value at the location (x, y) is obtained as follows:

$$Y_{x,y} = \max_{k \in [1,N]} \left(\mathbb{1}_{\{\|(x-\hat{c}_x^k)^2+(y-\hat{c}_y^k)^2\|_2 \le radius\}} \exp^{-\frac{(x-\hat{c}_x^k)^2+(y-\hat{c}_y^k)^2}{2\delta^2}} \right).$$

Here, $\mathbb{1}_{\{\|(x-\hat{c}_x^k)^2+(y-\hat{c}_y^k)^2\|_2 \le radius\}}$ is the indicator function, which is 1 if the distance between position (x, y) and the object location $(\hat{c}_x^k, \hat{c}_y^k)$ is less than or equal to the radius, and 0 otherwise. N stands for the number of heads, and δ is the standard deviation related to the radius. Specifically, $\delta = \frac{radius}{2}$.

The heatmap's loss function is the pixel-wise focal loss [16]:

$$L_{heat} = -\frac{1}{N} \sum_{xy} \begin{cases} (1 - \widetilde{Y_{x,y}})^{\alpha} \log\left(\widetilde{Y_{x,y}}\right) & Y_{x,y} = 1 \\ (1 - Y_{x,y})^{\beta} (\widetilde{Y_{x,y}})^{\alpha} \log(1 - \widetilde{Y_{x,y}}) & otherwise \end{cases} \tag{1}$$

where α and β are predetermined parameters, and $\widetilde{Y}$ is the predicted heatmap.

3.1.2. Box Size Branch and Offset Branch

The width and height of the target box are regressed at each position by the box size branch. The offset branch, meanwhile, aims to address the discretization issue brought on by the output stride. We denote box size branch output $\widetilde{S}$ and offset branch output $\widetilde{O}$. The dimensions of these two outputs both are $2 \times H \times W$. For each ground truth box $\mathbf{b}^k =$

$(x_1^k, y_1^k, x_2^k, y_2^k)$, its box size is $\mathbf{s}^k = (x_2^k - x_1^k, y_2^k - y_1^k)$ and the offset is $\mathbf{o}^k = (c_x^k - \hat{c}_x^k, c_y^k - \hat{c}_y^k)$. Then, we use l_1 losses for the box size branch and offset branch as follows:

$$L_{size} = \sum_{k=1}^{N} \|\mathbf{b}^k - \widetilde{\mathbf{b}}^k\|_1 \tag{2}$$

$$L_{off} = \sum_{k=1}^{N} \|\mathbf{o}^k - \widetilde{\mathbf{o}}^k\|_1 \tag{3}$$

where $\widetilde{\mathbf{b}}^k$ and $\widetilde{\mathbf{o}}^k$ denote the predicted box size and offset at the corresponding location, respectively.

3.1.3. Loss Function

The following is a definition of CRNet loss:

$$L = \lambda_1 L_{heat} + \lambda_2 L_{size} + \lambda_3 L_{off} \tag{4}$$

where the three hyper-parameters λ_1, λ_2, and λ_3 are used to balance the weights of three branches.

3.2. Global Context Refinement Module

The standard Non-local (NL) [30] is responsible for reinforcing the features of the query location by assembling features from other locations. The SNL [12] block is the simplified version of NL, which aims to compute a global attention map and shares this attention map for all query positions. SNL has a lower computation cost than and similar performance to NL. In Figure 2a, the detailed structure of the SNL block is illustrated. We denote $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{H \cdot W}$ as a feature map whose height is H and width is W. $\mathbf{x}$ and $\mathbf{y}$ represent the input and output of the SNL, respectively. The process of SNL block can be formulated as follows:

$$\mathbf{y}_i = \mathbf{x}_i + W_v \sum_{j=1}^{H \cdot W} \frac{\exp(W_k \mathbf{x}_j)}{\sum\limits_{k=1}^{H \cdot W} \exp(W_k \mathbf{x}_m)} \mathbf{x}_j \tag{5}$$

where i is the query position's index and j traverses all possible positions. Furthermore, W_v and W_k denote linear transformation matrices, which are both 1 × 1 conv.

When a feature map enters the SNL, the SNL first reduces the number of channels to 1 before using softmax for the global system to obtain a response map. The matrix multiplication for the features of the original input , which is the process of modeling long-range dependencies between each pair of points, is then performed using the response graph. Additionally, the multiplication results are added to the original input element-wise, resembling a residual structure. The reason why SNL can generate the response map with only one channel is that [12] discovered that the response maps of different query positions in NL are similar, so we only need to learn one to replace the others, thus greatly reducing the computational complexity.

Lin et al. [33] point out that, in the deep ConvNet, high-resolution maps have low-level features that hurt their representational capacity, while low-resolution maps have stronger semantics but coarser spatial representation ability. We rely on a certain architecture, the global context refinement module (GRM), which refines the global context via a top-down pathway and lateral connections in order to obtain a global context with strong representational ability and precise spatial positioning. Figure 2b depicts the basic block architecture of the GRM.

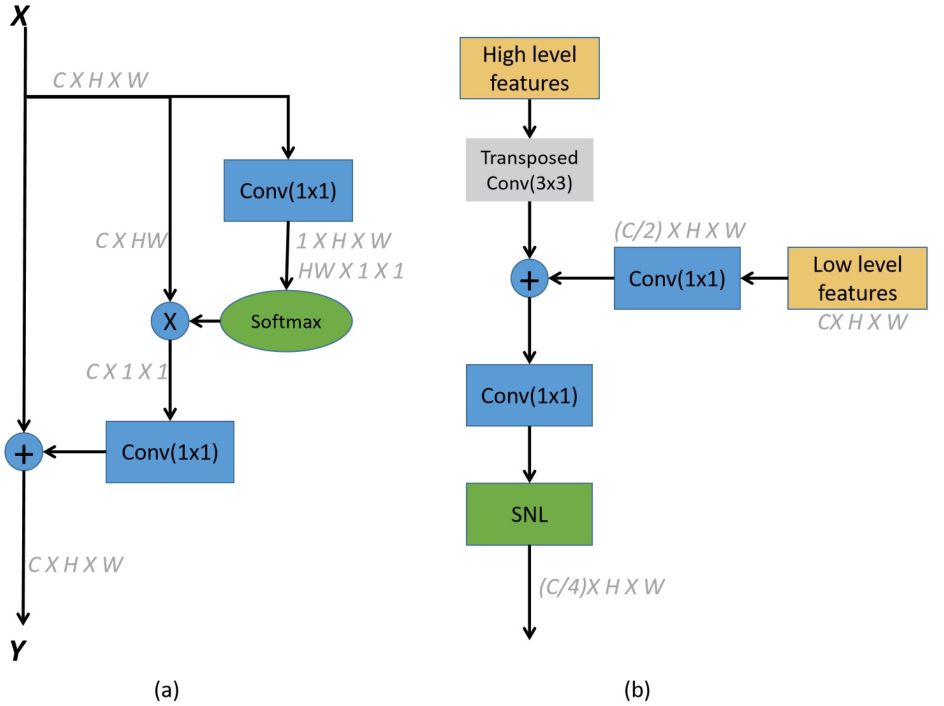

Figure 2. (**a**) SNL block; (**b**) GRM block; $C \times H \times W$ represents a feature map with channel number C, height H, and width W. Broadcast element-wise addition is denoted by $\oplus$, and matrix multiplication is denoted by $\otimes$.

All the low-level features go through a 1×1 convolutional layer to shrink the channel's dimensions. On account of the limited capability of the bilinear upsampling, our model applies a 3×3 transposed convolution to upsample the resolution of high-level features by a factor of 2 ($kernel = 4, stride = 2$). Then, using element-wise addition, the low-level feature map with reduced dimensions is combined with the corresponding upsampled feature map. For fused features, although parameters of SNL are much less than those of NL, we further reduce the channel dimensions through a 1×1 convolution for less computational complexity. Finally, the output feature map of 1×1 is used as the input of the SNL to obtain the refined global context by building long-range dependencies between similar pixels. We denote $M_i(i = 2,3,4)$ as the output as the GRM block which merges feature maps of S_i and $S_{i+1}(i = 2,3,4)$. In particular, M_5 is denoted as the first GRM block, which is only related to S_5 and has no upsampling or element-wise addition operator.

3.3. Local Context Refinement Module

Due to the perspective principle, the sizes of heads are different in pictures, even though the actual size difference is small. This situation of objects with different scales give rise to different sizes of local context ranges. A partial remedy is to use multi-scale convolution and apply larger kernel convolutions such as 5×5 or 7×7 to expand receptive field, which, however, requires additional memory and computation cost. We advocate instead the use of dilated convolution, inspired of impressive work in image segmentation [35–37]. The process of constructing dilated convolution consists of inserting zeros between each pixel in the convolutional kernel.

There are three main advantages of dilated convolution: (1) For a dilated convolution kernel with size $k \times k$ and dilation rate r, the size of its receptive field is $(k-1) \cdot (r-1)$ larger than that of a conventional convolution with the same kernel size. (2) If the stride of a dilated convolution is 1, it can maintain the resolution of input feature maps. (3) The dilated convolution matches the calculation amount of a standard convolution with the same kernel size. For instance, a dilated convolution whose kernel size is $k = 3$ and dilation rate is $r = 2$ has the same size receptive field as a standard convolution whose kernel size is 5×5, but without any change in Flops as a 3×3 convolution. In particular, in standard convolution, the dilation rate $r = 1$. As shown in Figure 3, we apply multi-scale dilation convolution with four branches with rates $(r = 6, 12, 18, 24)$. Each of them has $padding = r$ and $stride = 1$ to maintain the resolution of the input feature map. Additionally, we fuse them by adding their feature maps point-wise for refining the local context. After that, we use a residual architecture to prevent difficult training. Eventually, the feature map of LRM is passed through three parallel heads.

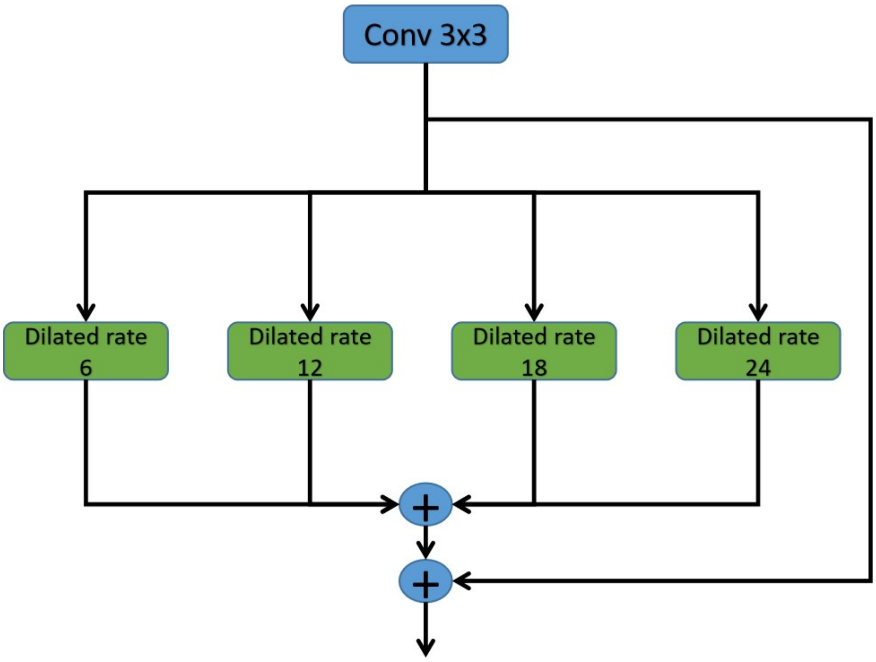

Figure 3. Illustration of LRM.

4. Experiments and Analysis

We detail the specifics of our experiments' execution in this section and examine the results. Datasets and evaluation metrics are first introduced. Next, we go into detail about how the experiments are conducted. We then use ablation studies to assess the efficiency of our method. Finally, our network is evaluated against the most advanced detectors on the public dataset.

4.1. Dataset and Evaluation Metrics

We conduct extensive research on the Brainwash [38] and HollywoodHeads [10] head detection datasets. The Brainwash dataset contains 82,906 labeled head boxes across 11,917 images. There are 10,917 total images in the training set, 500 images in the verification set, and 500 images in the test set. It is a large dataset obtained from crowd scenes utilizing video footage at a fixed interval. The HollywoodHeads dataset is the largest dataset currently used for head detection. It consists of 224,740 images from 21 Hollywood movies, with 369,846 heads in total. The training, validation, and test sets contain 216,719, 6719, and 1302 annotated images, respectively.

The evaluation metric adopted in testing is the standard average precision (AP). Furthermore, the IoU threshold is 0.5.

4.2. Implementation Details

Pre-trained weights from ImageNet are utilized as the model's initial parameters, and we adopt Adam as the optimizer, setting the initial learning rate to 3×10^{-4}. The training batchsize is set to 32 (on 2 Nividia 3060 GPUs). Furthermore, we set the max training epochs to 65 and 20 for the Brainwash and HollywoodHeads datasets, respectively.

We limit the input size of images to 512 × 512 during the training phase. For the data augmentation, random horizontal flips, random scaling from 0.6 to 1.3, random cropping, and color jittering are used in this implementation.

At inference time, all test samples are resized to 512 × 512. Following [18], we first extract the peaks of the heatmaps whose responses are greater than or equal to the connected pixels around them by using a 3 × 3 max pooling operation. We then keep the top 100 peaks as the head center points. The number of points is further filtered using a confidence threshold, which we set to 0.05 in the test phase. Let P be one of the detected head center points with the heatmap location (x_p, y_p). The detected score of P is $\widetilde{Y}_{x_p,y_p}$, and its prediction bounding box can be produced as follows:

$$(x_p + \Delta x_p - \frac{w_p}{2}, y_p + \Delta y_p - \frac{h_p}{2}, x_p + \Delta x_p + \frac{w_p}{2}, y_p + \Delta y_p + \frac{h_p}{2})$$

where $(w_p, h_p) = \widetilde{S}_{x_p,y_p}$ and $(\Delta x_p, \Delta y_p) = \widetilde{O}_{x_p,y_p}$.

4.3. Ablation Analysis

Considering that the output form and loss function of our model are the same as CenterNet [18], and the model architecture is similar, we adopt CenterNet as the baseline. Following the original CenterNet [18], we set the three hyper-parameters λ_1, λ_2, and λ_3 to 1, 0.1, and 1, respectively, in Equation (4) and set α and β to 1 and 2 in Equation (1). For fair comparison, we employ ResNet50 instead of Dla34 [39] as the backbone relative to the standard CenterNet. In particular, the upsampling operator of CenterNet(ResNet50) is three deconvolution layers. We use the same settings to achieve a fair comparison during the training stage. For instance, the initialization of the parameters is the same for all models, and the input size is 512 × 512.

Table 1 presents the ablation experiment's outcomes. We can see that our GRM and LRM can each outperform the baseline model in terms of AP by 2.29% and 1.04%, respectively. On the Brainwash test dataset, using LRM and GRM simultaneously can even improve performance by 2.71% over the benchmark model. Figure 4 displays the precision recall (PR) curves of the ablation analysis of CRNet at IoU 0.5 on the Brainwash dataset. These results show that LRM and GRM can effectively improve the performance of the head detector.

Table 1. The ablation study of CRNet on the Brainwash test dataset.

Model Name	**Methodology**		**AP (%)**
	GRM	**LRM**	
CenterNet (ResNet50)			91.18
CRNet	✓		93.47
CRNet		✓	92.22
CRNet (proposed)	✓	✓	93.89

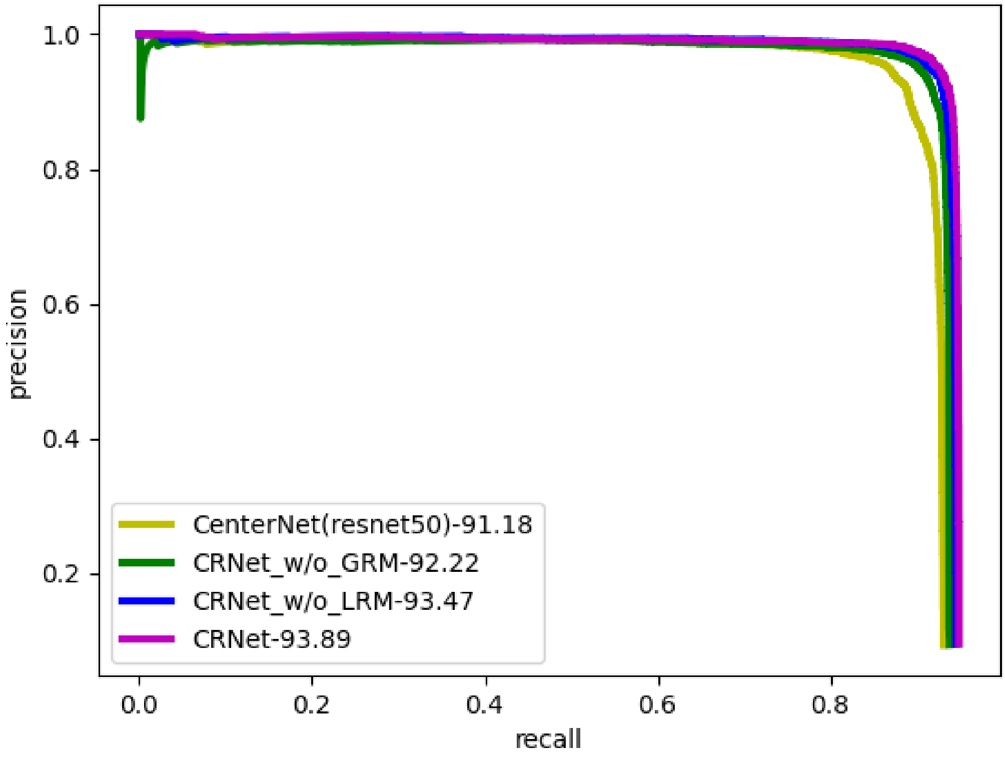

Figure 4. The PR curves of ablation analysis at IoU 0.5 on the Brainwash dataset.

Additionally, we further compare the influence of the number of GRM blocks used in CRNet by successively adding GRM M_5, M_4, M_3, and M_2 on the CRNet without LRM. As indicated in Table 2, GRM is applied to all four stages for the best outcome.

Table 2. The influence of the number of GRM blocks on the Brainwash test dataset.

Position of GRM	AP (%)
None	91.18
S_5	91.45
$S_5 + S_4$	92.36
$S_5 + S_4 + S_3$	92.82
$S_5 + S_4 + S_3 + S_2$	93.47

Another observation from Table 2 is that the absolute improvement of adding GRM on S_5 is the smallest. One possible explanation is that although the representation ability is strong in the global context at the high stage, the spatial representation ability is weak due to the large receptive field.

Furthermore, we display qualitative results of GRM on the Brainwash test set. We exploit Grad-CAM [40] to visualize the class activation map (CAM)[41] for SNL blocks from different GRMs in Figure 5. Input images and the CAMs of the M_5, M_4, M_3, and M_2 are displayed in order from left to right. The CAM will highlight the area that has more influence on the result. The greater the contribution of the area to the final result, the more the area will be highlighted. It can be clearly seen that the process of the global context of images is refined step by step. In Figure 5, from left to right, almost all head center points become clearer and more accurate. The ability of GRM to obtain the global context and gradually refine the global context by combining features from various resolutions is demonstrated by this process. It can be determined from the last row in the figure that even the completely occluded head (left bottom) can be recalled through the refined global context, which is nearly impossible using the local context.

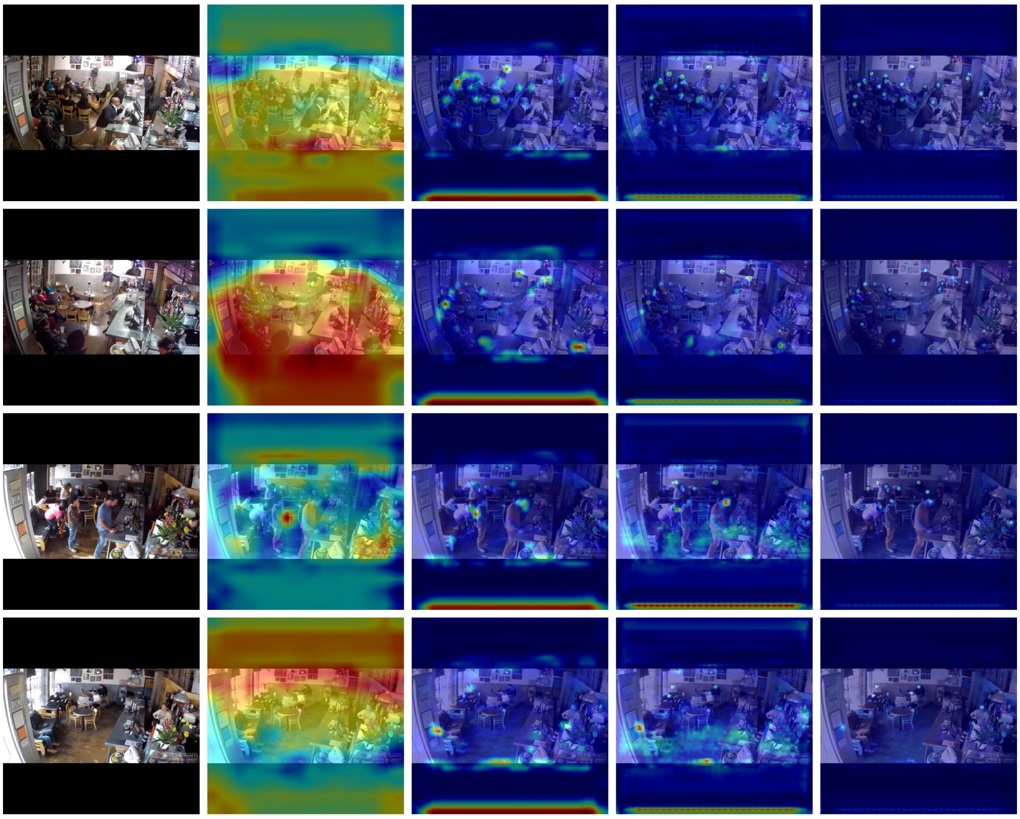

Figure 5. Input images and the CAMs of the M_5, M_4, M_3, and M_2 are displayed in order from left to right.

We also experiment with various configurations of dilation rates on our model with GRM to find an optimal combination of LRM. The results in Table 3 show that multi-scale dilated convolution can have a positive effects in the results, and the configuration of $\{6, 12, 18, 24\}$ is the best combination of dilation rates for our experiments.

Table 3. The results of different configurations of dilation rates on the Brainwash test dataset.

Configurations of Dilation Rates	AP (%)
{6}	93.46
{6, 12}	93.67
{6, 12 ,18}	93.84
{6, 12, 18, 24}	93.89
{4, 8, 12 , 16}	93.58

4.4. Performance Evaluation

According to the ablation studies above, we adopt GRM on the outputs of all stages and a multi-scale dilated convolution with dilatation rates of $\{6, 12, 18, 24\}$. Finally, we evaluate the performance of our detector against several other representative detectors, including SSD, Faster-RCNN, TINY [42], and the existing methods for head detection, such as HeadNet [13] and SANM [32]. Table 4 offers a summary of the results from the Brainwash dataset. We provide some examples in Figure 6. These examples indicate that our model has good performance in densely populated scenes and small target detection.

Figure 7 compares CRNet with other head detectors that show good performance by plotting the PR curves at IoU 0.5 on the Brainwash dataset. Our method shows a better performance than other detectors, which proves the superiority of our proposed model. A competitive performance is obtained when we test our model on HollywoodHeads as well. In Table 5, the outcomes for HollywoodHeads are presented. Figure 8 presents the qualitative results of our method on the HollywoodHeads dataset.

Table 4. The results on the Brainwash dataset.

Model	**AP (%)**
SSD [8]	74.1
HSFA2Net [43]	89.2
TINY [42]	89.3
E2PD [38]	82.1
HeadNet [13]	91.3
Faster-RCNN [7]	91.9
SANM [32]	93.12
CRNet (ours)	93.89

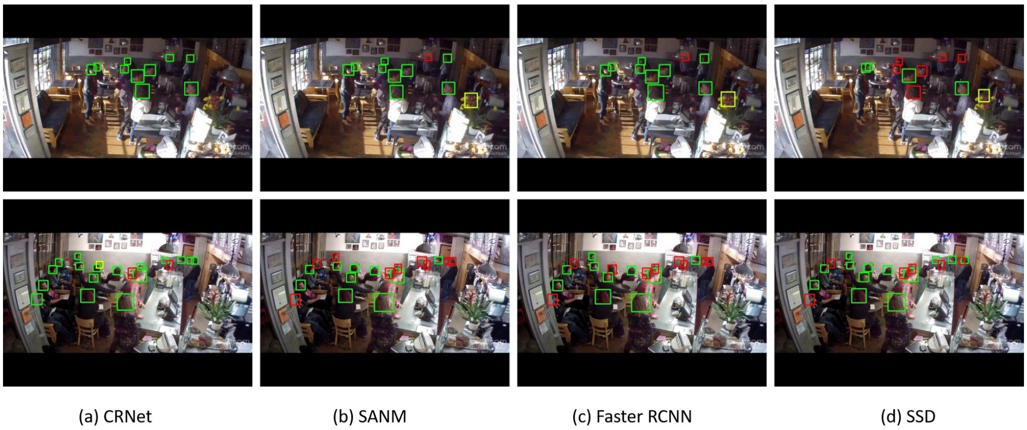

Figure 6. The qualitative results of our method on the Brainwash dataset. We present the results of four different methods. Red boxes indicate the missed heads, while yellow boxes represent samples for error detection, and green boxes are detection results that match the ground truth.

Table 5. The results on the HollywoodHeads dataset.

Model	**AP (%)**
CACNN [10]	72.7
MSRF [22]	81.0
SSD [8]	81.1
TINY [42]	81.3
HeadNet [13]	83.0
CRNet (ours)	83.4

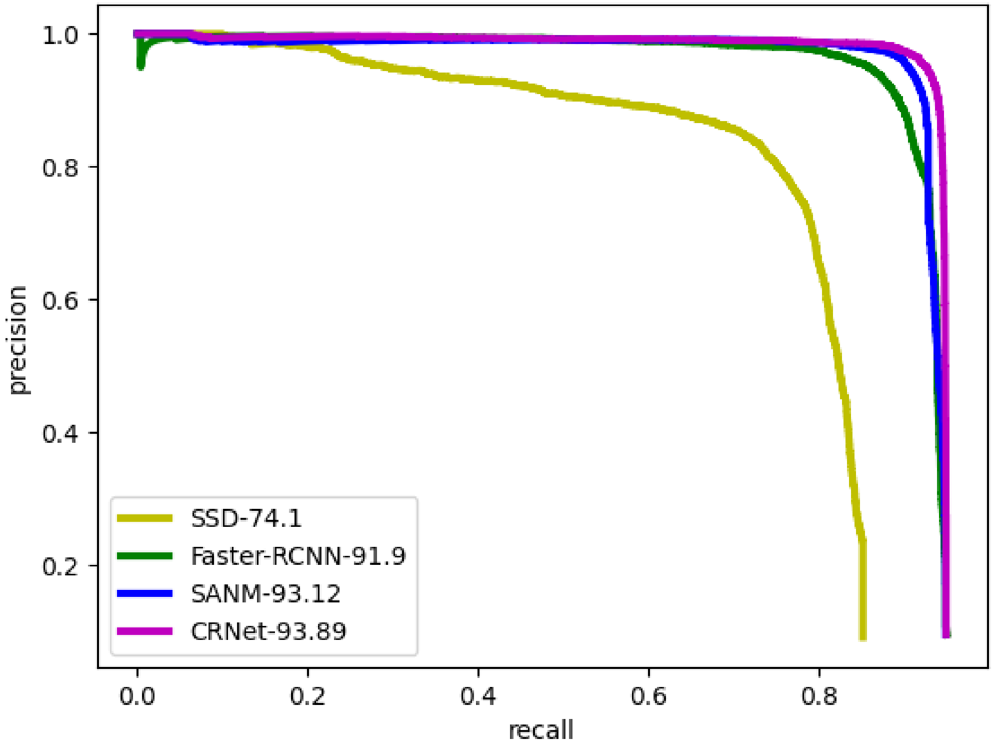

Figure 7. The PR curves of our model and other methods at IoU 0.5 on the Brainwash dataset.

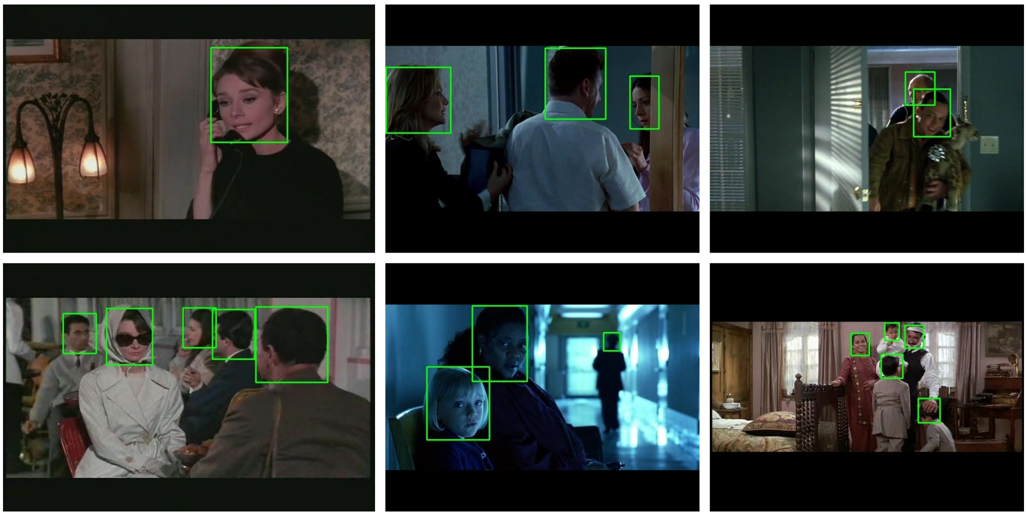

Figure 8. Examples of CRNet's head detection results on the HollywoodHeads dataset.

In terms of running time, we use a Ubuntu platform with one NVIDIA 3060 GPU and an Intel Xeon CPU E5-2697 v2 @ 2.70GHz to test the running speeds of different algorithms. At the same time, we also test the effect of LRM and GRM on the running speed of our model. As shown in Table 6, our model achieves a good balance between precision and speed. Furthermore, as expected, LRM and GRM do not have a very large negative impact on the operating speed.

Table 6. The results of running time tests.

Model	Backbone	AP (%)	FPS
SSD	VGG16	74.1	56.0
Faster RCNN	ResNet50	91.9	15.4
SANM	ResNet50	93.12	15.2
CRNet w/o GRM	ResNet50	92.22	48.5
CRNet w/o LRM	ResNet50	93.47	49.3
CRNet (proposed)	ResNet50	93.89	47.7

5. Conclusions and Future Work

Although head detection is a fundamental and important task, the current approaches fall short in terms of their ability to extract useful local and global contexts. We suggest CRNet, which is a one-stage detector for head detection, as a remedy for this issue. CRNet identifies heads via their center points in the images, without any anchor design, avoiding the need for prior statistical analysis. Because we associate heads with their center points, the similarity between points replaces the similarity between head proposals, and the computational costs become irrelevant to the number of targets. In order to extract the global context better and improve the classification ability, we introduce the global context refinement module (GRM). The GRM establishes long-range dependencies between pixels through SNL. SNL is much easier to optimize than the previous approach to obtaining global contexts through the use of repeated convolution layers [12]. In addition, we enhance the global context by fusing output feature maps of different GRMs. Furthermore, a multi-scale dilated convolution layer is also used to refine the local context around the center point of the head. The computational complexity of the dilated convolution with a dilation rate greater than 1 is less than that of a conventional convolution with the same receptive field. The results of our experiments demonstrate that our proposed method achieves state-of-the-art performance on two public datasets. The running time test proves that our method achieves a good balance in terms of speed and accuracy. In future work, we intend to expand our framework to other visual fields, such as human crowd counting and face detection. At the same time, inspired by the latest progress in deep learning technology, in future work, we will explore the differences in our models' performance caused by different backbones, especially ViT backbones [44].

Author Contributions: Conceptualization, Z.Y.; methodology, J.H.; software, J.H.; validation, J.H. and Z.Y.; formal analysis, J.H.; investigation, J.H.; resources, Z.Y.; data curation, J.H.; writing—original draft preparation, J.H.; writing—review and editing, J.H. and Z.Y.; visualization, J.H.; supervision, Z.Y.; project administration, Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Brainwash Dataset can be found at https://exhibits.stanford.edu/data (accessed on 22 November 2022), and the HollywoodHeads Dataset can be downloaded from https://www.robots.ox.ac.uk/~vgg/software/headmview/ (accessed on 22 November 2022).

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

Abbreviations

The following abbreviations are used in this manuscript:

CRN	Context refinement network
GRM	Global context refinement module
LRM	Local context refinement module
NL	Non-local
SNL	Simplified Non-local
AP	Average precision
PR	Precision recall
CAM	Class activation map
RPN	Region proposal network

References

1. Hu, J.; Lu, J.; Tan, Y.P. Deep metric learning for visual tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *26*, 2056–2068. [CrossRef]
2. Senior, A.; Pankanti, S.; Hampapur, A.; Brown, L.; Tian, Y.L.; Ekin, A.; Connell, J.; Shu, C.F.; Lu, M. Enabling video privacy through computer vision. *IEEE Secur. Priv.* **2005**, *3*, 50–57. [CrossRef]
3. Shami, M.B.; Maqbool, S.; Sajid, H.; Ayaz, Y.; Cheung, S.C.S. People counting in dense crowd images using sparse head detections. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 2627–2636. [CrossRef]
4. Lu, J.; Liong, V.E.; Zhou, X.; Zhou, J. Learning compact binary face descriptor for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 2041–2056. [CrossRef] [PubMed]
5. Lu, J.; Wang, G.; Zhou, J. Simultaneous feature and dictionary learning for image set based face recognition. *IEEE Trans. Image Process.* **2017**, *26*, 4042–4054. [CrossRef] [PubMed]
6. Yi, S.; Li, H.; Wang, X. Understanding pedestrian behaviors from stationary crowd groups. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015 ; pp. 3488–3496.
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. Available online: https://arxiv.org/abs/1506.01497 (accessed on 22 November 2022).
8. 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: Berlin/Heidelberg, Germany, 2016, pp. 21–37.
9. 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.
10. Vu, T.H.; Osokin, A.; Laptev, I. Context-aware CNNs for person head detection. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2893–2901.
11. Sempau, J.; Wilderman, S.J.; Bielajew, A.F. DPM, a fast, accurate Monte Carlo code optimized for photon and electron radiotherapy treatment planning dose calculations. *Phys. Med. Biol.* **2000**, *45*, 2263. [CrossRef] [PubMed]
12. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
13. Li, W.; Li, H.; Wu, Q.; Meng, F.; Xu, L.; Ngan, K.N. Headnet: An end-to-end adaptive relational network for head detection. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 482–494. [CrossRef]
14. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212.
15. Zhang, S.; Wen, L.; Lei, Z.; Li, S.Z. RefineDet++: Single-shot refinement neural network for object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 674–687. [CrossRef]
16. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
17. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
18. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
19. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
20. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
21. Bell, S.; Zitnick, C.L.; Bala, K.; Girshick, R. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2874–2883.

22. Wang, Y.; Yin, Y.; Wu, W.; Sun, S.; Wang, X. Robust person head detection based on multi-scale representation fusion of deep convolution neural network. In Proceedings of the 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO), Macau, China, 5–8 December 2017; pp. 296–301.
23. Zhou, T.; Li, J.; Wang, S.; Tao, R.; Shen, J. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Trans. Image Process.* **2020**, *29*, 8326–8338. [CrossRef] [PubMed]
24. Zhou, T.; Li, L.; Bredell, G.; Li, J.; Unkelbach, J.; Konukoglu, E. Volumetric memory network for interactive medical image segmentation. *Med. Image Anal.* **2023**, *83*, 102599. [CrossRef] [PubMed]
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: https://arxiv.org/abs/1706.03762 (accessed on 22 November 2022).
26. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning. PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363.
27. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
28. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
29. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. Ocnet: Object context network for scene parsing. *arXiv* **2018**, arXiv:1809.00916.
30. 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, 18–22 June 2018; pp. 7794–7803.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
32. Zhang, J.; Liu, Y.; Li, R.; Dou, Y. End-to-end Spatial Attention Network with Feature Mimicking for Head Detection. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 199–206.
33. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
34. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9759–9768.
35. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
36. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
37. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
38. Stewart, R.; Andriluka, M.; Ng, A.Y. End-to-end people detection in crowded scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2325–2333.
39. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep layer aggregation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2403–2412.
40. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
41. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
42. Hu, P.; Ramanan, D. Finding tiny faces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 951–959.
43. Shen, W.; Qin, P.; Zeng, J. An indoor crowd detection network framework based on feature aggregation module and hybrid attention selection module. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
44. Beal, J.; Kim, E.; Tzeng, E.; Park, D.H.; Zhai, A.; Kislyuk, D. Toward transformer-based object detection. *arXiv* **2020**, arXiv:2012.09958.