



Article

Global Contextual Dependency Network for Object Detection

Junda Li ^{1,2} , Chunxu Zhang ^{1,2} and Bo Yang ^{1,2,*}¹ College of Computer Science and Technology, Jilin University, Changchun 130012, China; ljd19@mails.jlu.edu.cn (J.L.); cxzhang19@mails.jlu.edu.cn (C.Z.)² Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

* Correspondence: ybo@jlu.edu.cn

Abstract: Current two-stage object detectors extract the local visual features of Regions of Interest (RoIs) for object recognition and bounding-box regression. However, only using local visual features will lose global contextual dependencies, which are helpful to recognize objects with featureless appearances and restrain false detections. To tackle the problem, a simple framework, named Global Contextual Dependency Network (GCDN), is presented to enhance the classification ability of two-stage detectors. Our GCDN mainly consists of two components, Context Representation Module (CRM) and Context Dependency Module (CDM). Specifically, a CRM is proposed to construct multi-scale context representations. With CRM, contextual information can be fully explored at different scales. Moreover, the CDM is designed to capture global contextual dependencies. Our GCDN includes multiple CDMs. Each CDM utilizes local Region of Interest (RoI) features and single-scale context representation to generate single-scale contextual RoI features via the attention mechanism. Finally, the contextual RoI features generated by parallel CDMs independently are combined with the original RoI features to help classification. Experiments on MS-COCO 2017 benchmark dataset show that our approach brings continuous improvements for two-stage detectors.

Keywords: object detection; global contextual dependency; multi-scale representations; attention mechanism



Citation: Li, J.; Zhang, C.; Yang, B. Global Contextual Dependency Network for Object Detection. *Future Internet* **2022**, *14*, 27. <https://doi.org/10.3390/fi14010027>

Academic Editor: Paolo Bellavista

Received: 15 December 2021

Accepted: 10 January 2022

Published: 13 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Object detection aims at locating and recognizing object instances from predefined object categories [1]. The significant progress of Convolutional Neural Networks (CNNs) [2,3] has brought excellent breakthroughs in object detection. In general, CNN-based detectors could be divided into two types, two-stage, and single-stage detectors. Our focus is on two-stage detectors. In current two-stage approaches, the Region of Interest (RoI) head extracts visual features of RoIs to predict specific categories and refine locations. While the theoretical receptive field of each RoI is large, the effective receptive field [4] remains limited, which makes the local visual features of RoIs lack global contextual dependencies.

In the physical world, visual objects have natural context dependencies relationships with the particular environment (i.e., background) and other related objects (i.e., foreground) [5]. For instance, as shown in Figure 1a, a person, tennis racket, and sports ball often appear together on the tennis court. The context dependencies are helpful to recognize objects with featureless appearances or restrain noisy detections [6]. As shown in Figure 1b, only part of the tennis racket appears in the image, and the lack of visual features makes the tennis racket unrecognizable (left). Taking into account the person and even the clothes, the tennis racket is correctly detected (right). On the other hand, as shown in Figure 1c, the background, traffic light, and clock are discriminative clues for eliminating the false detection of sports balls.

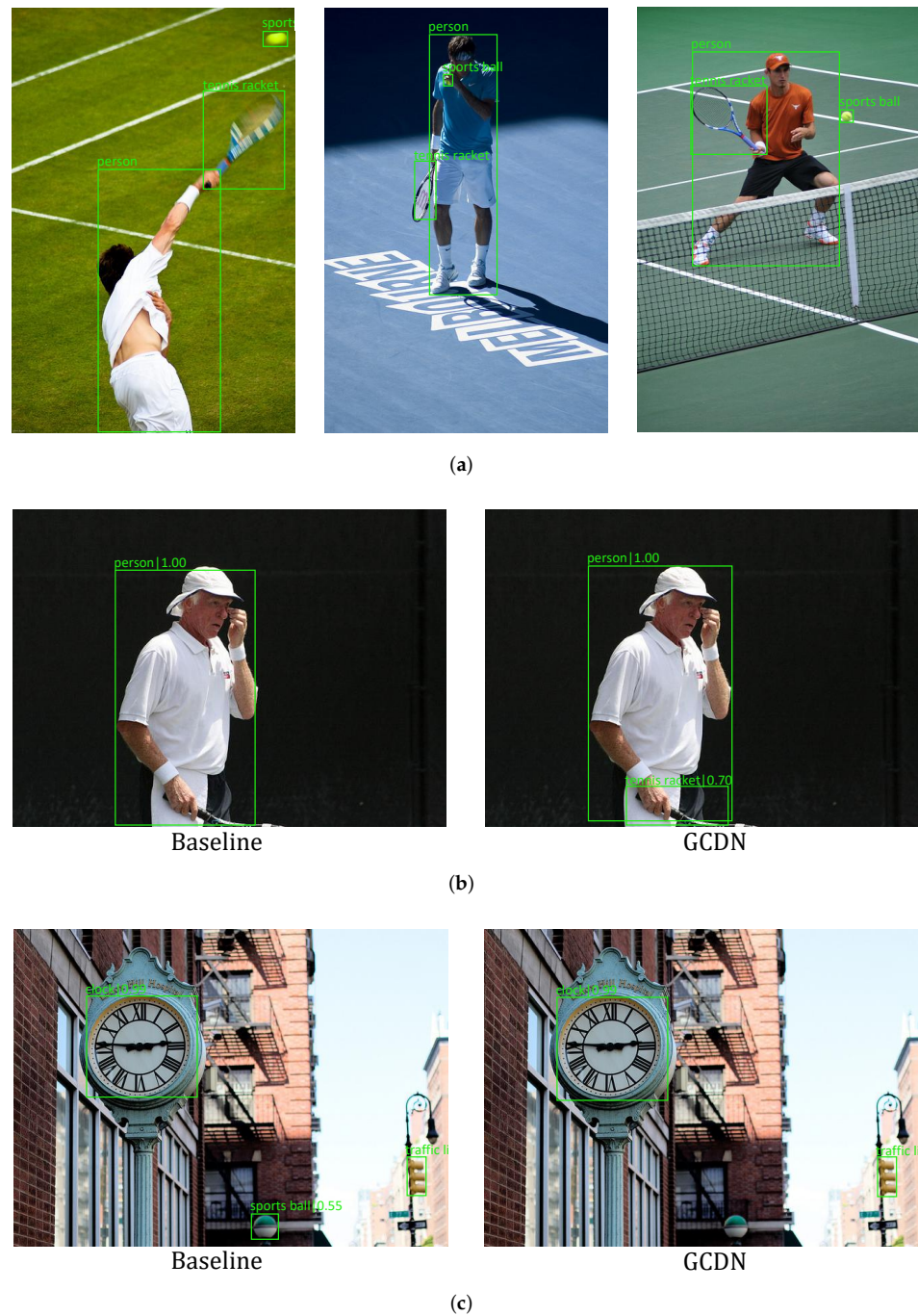


Figure 1. Examples and effects of the context dependencies. (a) Person, tennis racket and sports ball often appear together on the tennis court. (b) recognizing featureless objects. (c) restraining noisy detections.

Various methods have attempted to capture the context dependencies by modeling the visual and spatial relationships between RoIs [7,8]. However, the sampled RoIs may only cover a part of the image, leading to the omission of information in some regions. Global context refers to image-level context and is useful to capture global contextual dependencies [9,10]. In practice, simple use of global context (i.e., global average pooled context representation and RoI features are concatenated before recognition) leads to improvements. However, there are two challenges in efficiently utilizing the global context. On the one hand, the global context also contains noisy information, making us aware that a careful identification of global contextual dependencies contained in the global

context is desired. On the other hand, it's noted that the global context lacks explicit spatial information such as location and size. To close the gap, exploring the visual features in the global context from a spatial perspective is essential.

To achieve the goal, we present a simple Global Contextual Dependency Network (GCDN), which captures global contextual dependencies over local visual features to further enhance the local RoI feature representations. Considering that visual objects varies substantially in scale and are distributed in different locations, a Context Representation Module (CRM) is exploited to construct multi-scale context representations. Furthermore, to capture global contextual dependencies, we utilize the attention mechanism and design a Context Dependency Module (CDM). Our GCDN consists of multiple CDMs. Each CDM generates single-scale contextual RoI features based on the local RoI features and single-scale context representation via affinity computation and context aggregation. Multi-scale contextual RoI features generated by parallel CDMs independently are fused with the original RoI features to predict the labels of the RoIs.

Comprehensive experiments are performed to validate the effectiveness and generality of GCDN. Our GCDN improves 1.5% and 1.2% Average Precision (AP) on MS-COCO 2017 benchmark dataset [11] with ResNet-50 for Feature Pyramid Network (FPN) and Mask R-CNN, respectively. Ablation studies show that the CRM and CDM complement each other to improve the detection results.

Our contributions are summarized as follows:

- We present a novel Global Contextual Dependency Network (GCDN), as a plug-and-play component, to boost the classification ability of two-stage detectors;
- A Context Representation Module (CRM) is proposed to construct multi-scale context representations, and a Context Dependency Module (CDM) is designed to capture global contextual dependencies;
- Our proposed GCDN significantly improves detection performance and is easy to implement. Furthermore, we propose a lite version for little calculation.

2. Related Work

2.1. Object Detection

Deep learning technology learns feature representations end-to-end and has made extraordinary progress in object detection [12], semantic segmentation [13,14] and other vision applications [15,16]. In general, deep learning-based detectors are mainly divided into two types, single-stage (e.g., YOLO [17], SSD [18]) and two-stage detectors (e.g., Fast R-CNN [19], Faster R-CNN [20]). Two-stage approaches generally have slower speeds than single-stage approaches but have better detection performance [21].

As a classic two-stage detector, Faster R-CNN [20] designs a novel Region Proposal Network (RPN) to generate rectangular proposals and promotes the emergence of follow-up works [22,23]. For instance, Feature Pyramid Network (FPN) [22] addresses multi-scale problems using feature pyramid representations. Mask R-CNN [24] proposes an RoIAlign layer to align the visual features of objects exactly. These methods improve the quality of local visual features while neglecting the context dependencies. As a complement to these works, we concentrate on capturing context dependencies to further enhance the local RoI feature representations.

2.2. Context Dependency for Object Detection

Appropriate modeling of context dependencies is beneficial to object detection and recognition [25]. Current methods have explored exploiting context dependencies to improve performance via self-attention mechanism or recurrent neural networks [8,26–28]. Structure Inference Network (SIN) [29] uses Gated Recurrent Unit (GRU) to propagate messages among objects and the scene. Relation Network [7] introduces the self-attention mechanism [30] to model the context dependencies relationships between RoIs. Inside-Outside Net (ION) [31] exploits spatial recurrent neural networks to capture global contextual dependencies but leaves spatial relationships such as size out of consideration.

The context dependencies captured by the above methods involve visual and spatial relationships between RoIs. However, RoIs are sampled regions in the image, which may cause some other important regions to be ignored. On the other hand, the modeling of spatial relationships plays an important role in capturing context dependencies. In contrast, this work captures global contextual dependencies and integrates spatial relationships into visual relationships implicitly.

3. Global Contextual Dependency Network

This part starts from the Global Contextual Dependency Network (GCDN) framework overview (see Figure 2). Firstly, a Context Representation Module (CRM) is employed to construct multi-scale context representations to explore contextual information at different scales. Then, multiple Context Dependency Modules (CDMs) are designed to capture global contextual dependencies. Specifically, each CDM generates single-scale contextual RoI features for each scale context representation independently. Finally, the multi-scale contextual RoI features generated by parallel CDMs and the original RoI features are fused to predict specific categories. The details of the CRM and CDM are elaborated as follows.

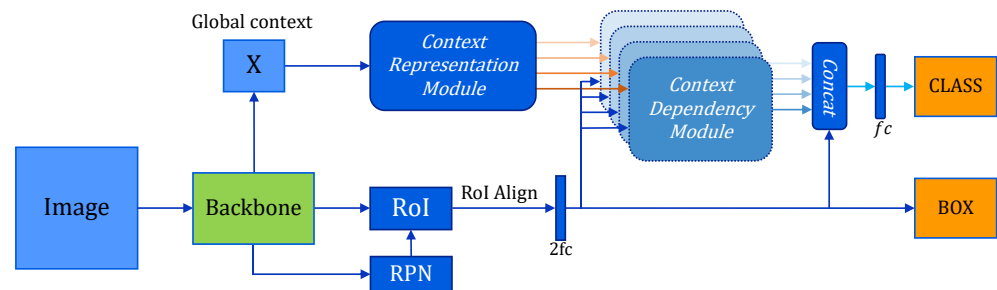


Figure 2. An overview of our Global Contextual Dependency Network (GCDN).

3.1. Context Representation Module

Visual objects appearing in an image often have various sizes and different locations. To integrate spatial information into visual features, the Context Representation Module (CRM) aims to construct multi-scale context representations.

As shown in Figure 2, the backbone CNN (e.g., ResNet) is firstly employed to extract visual features of the input image. Formally, the output convolutional feature map of CNN (e.g., the output of ResNet conv5) is denoted as global context $X \in \mathbb{R}^{C \times H \times W}$, where C represents the channel dimensionality, H and W represent the spatial height and width, respectively. Assuming that the stride between the input image and the global context is D (i.e., the downsample ratio of backbone CNN), then each position in X may represent a region containing $D \times D$ pixels in the input image. The single receptive field scale prevents us from fully exploring the contextual information.

To bridge the gap, the CRM generates multi-scale context representations based on X with S pyramid scales. Taking one scale s as an instance (see Figure 3), we firstly regard $s \times s$ positions in X as a region, and then aggregate its contents into a position by average pooling. By this way, a single-scale context representation $X^s \in \mathbb{R}^{C \times \lfloor H/s \rfloor \times \lfloor W/s \rfloor}$ is obtained, and each position in X^s may represent a region containing $Ds \times Ds$ pixels in the input image. The other scales are processed similarly. With multi-scale context representations, the visual features of objects with large size variations are captured exactly.

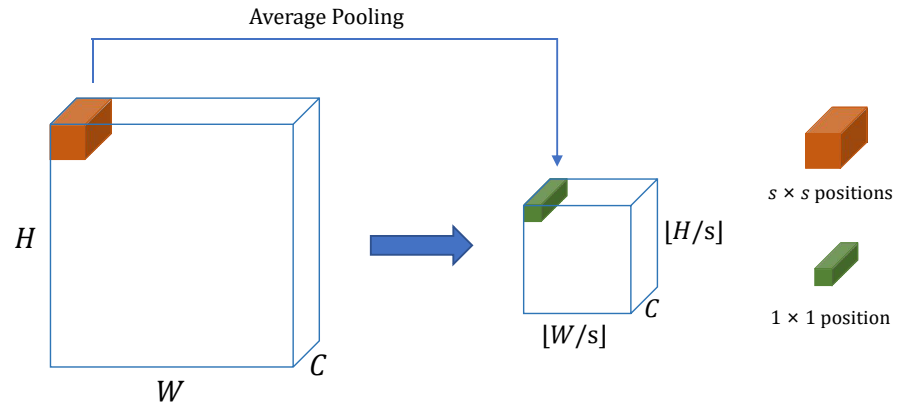


Figure 3. An illustration of the context representation operation (better viewed in color).

3.2. Context Dependency Module

As shown in Figures 2 and 4, GCDN includes multiple Context Dependency Modules (CDMs) organized in parallel. Each CDM takes RoI features and single-scale context representation as input and generates single-scale contextual RoI features by affinity computation and context aggregation. The multi-scale contextual RoI features generated by parallel CDMs and the original RoI features are concatenated to perform classification.

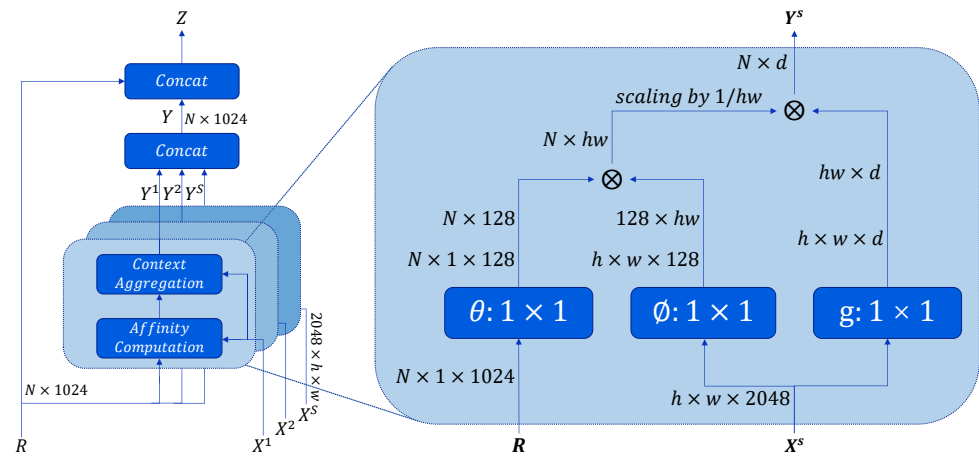


Figure 4. The pipeline of multiple Context Dependency Modules (CDMs).

Mathematically, let $R = \{r_i\}_{i=1}^N$ and $X^s = \{x_j^s\}_{j=1}^{h \times w}$ denote the RoI features and the single-scale context representation with pyramid scale s , respectively. r_i is the representation vector of i^{th} RoI, and N is the number of RoIs. x_j^s is the representation vector of j^{th} position, and $h \times w$ denotes the total number of positions in X^s .

3.2.1. Affinity Computation

The computation of affinity is given as follows:

$$\omega_{ij} = \frac{1}{C(X^s)} f_{\theta}(r_i)^T f_{\phi}(x_j^s), \quad (1)$$

where ω_{ij} denotes the impact of j^{th} position on i^{th} RoI. $f_{\theta}(\cdot)$ and $f_{\phi}(\cdot)$ are the query transform function and the key transform function respectively (implemented as 1×1 convolution). To reduce computational cost, θ and ϕ are dimensionality-reduction layers. $C(X^s)$ is a normalization factor whose value is $h \times w$.

3.2.2. Context Aggregation

After computing affinity between each RoI and each position in X^s , the single-scale contextual RoI features are reallocated according to the affinity and the context representation:

$$y_i^s = \sum_{j=1}^{h \times w} w_{ij} g(x_j^s), \quad (2)$$

where y_i^s denotes the single-scale contextual feature of i^{th} RoI, $g(\cdot)$ is the value transform function (implemented as 1×1 convolution). Let $Y^s = \{y_i^s\}_{i=1}^N$ denote the single-scale contextual RoI features.

3.2.3. Feature Fusion

As shown in Figure 4, the multiple Y^s generated from parallel CDMs independently are concatenated as the multi-scale contextual RoI features Y :

$$Y = \text{Concat}(Y^1, \dots, Y^S), \quad (3)$$

where Y has the same channel dimension as R (i.e., the number of channel dimensionality d shown in Figure 4 is related to the number of CDMs).

Concatenated with the contextual RoI features, the original RoI features are enhanced to form the final RoI features Z :

$$Z = \text{Concat}(R, Y). \quad (4)$$

The enhanced RoI features Z firstly pass through a fc head which includes one linear layer with ReLU activation (keep the same channel dimension as R) and are then used to predict the labels.

4. Experiments

Our GCDN could be used as a plug-and-play component for two-stage detectors. To verify the effectiveness and generality, we experiment on the MS-COCO 2017 benchmark dataset [11], which contains rich contextual information. MS-COCO 2017 contains 80 object categories of various sizes. All models are trained on the 118K training images and evaluated on the 5K validation (*val*) images. The Average Precision (AP) with IoU thresholds from 0.5 to 0.95 and 0.05 interval is taken as the evaluation metric. Two popular detectors, Feature Pyramid Network (FPN) [22] and Mask R-CNN [24], are taken as our baselines. Both two baselines have made significant improvements in extracting local visual features, thus better reflecting the role of GCDN.

4.1. Implementation Details

Our approach is based on MMDetection [32] codebase and all models follow the given settings. The short side of the input image is resized to 800 and the long side is no more than 1333. Random horizontal flipping is used as the only data augmentation operation during training. ImageNet [33] pre-trained ResNet-50 and ResNet-101 [3] are taken as backbone networks (ResNet-50 is taken unless specified otherwise). All models are trained with a batch size of 8 images for 12 epochs on 4 GPUs. The learning rate is initialized with 0.01 and decreased by 0.1 after 8 and 11 epochs, respectively. SGD with 0.9 momentum and 0.0001 weight decay is used as the optimizer. Pyramid scales $\{1, 2, 3, 6\}$ are adopted unless specified otherwise.

4.2. Comparisons with Baselines

The overall performance of our GCDN with different baselines and different backbones on MS-COCO 2017 *val* are shown in Table 1. Our model achieves continuous gains with all baselines. In particular, our GCDN improves 1.5% and 1.2% AP with ResNet-50-FPN for

FPN and Mask R-CNN, respectively. Furthermore, it is observed that the improvements for medium objects (AP_m) are significant. We conjecture that the visual features of medium objects are slightly insufficient and their existences are more dependent on the context. As powerful complementary information, the global contextual dependencies make the original RoI features more discriminative.

Table 1. Comparisons with baselines on MS-COCO 2017 *val*.

Backbone	Method	GCDN	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
ResNet-50-FPN	FPN		37.4	58.1	40.8	21.4	40.8	48.5
ResNet-50-FPN	FPN	✓	38.9	60.3	41.9	22.7	42.6	49.8
ResNet-50-FPN	Mask R-CNN		38.2	58.9	41.5	22.4	41.6	49.7
ResNet-50-FPN	Mask R-CNN	✓	39.4	60.5	42.8	23.0	43.5	50.6
ResNet-101-FPN	FPN		39.6	60.6	43.3	22.7	43.6	52.2
ResNet-101-FPN	FPN	✓	40.3	61.5	43.8	23.8	44.5	53.0
ResNet-101-FPN	Mask R-CNN		40.2	60.4	44.1	22.9	44.1	53.3
ResNet-101-FPN	Mask R-CNN	✓	41.0	62.1	44.6	24.0	45.1	53.7

4.3. Ablation Studies

In this section, two ablation experiments are conducted to analyze the presented modules.

4.3.1. Context Operations

This part investigates the effects of our CRM and CDM in our GCDN framework. Specifically, the CRM constructs multi-scale context representations, and the CDM aggregates context representation by attention mechanism. For fair comparisons, the global average pooling operation (denoted as “GAP”) is also utilized to aggregate context representation. Table 2 shows the experimental results. As aforementioned, just simply concatenating global average pooled context representation with RoI features before recognition improves 0.4% AP. In addition, a careful identification using the attention mechanism brings another 0.7% AP improvement. Finally, the CDM gives 0.4% AP improvement. The results verify that the CDM effectively identifies the global contextual dependencies and the CRM successfully captures the contextual information at different scales.

Table 2. Effects of different context operations on MS-COCO 2017 *val*.

Method	CRM	GAP	CDM	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
FPN				37.4	58.1	40.8	21.4	40.8	48.5
FPN		✓		37.8	59.1	41.0	21.9	41.4	48.3
FPN			✓	38.5	59.9	42.0	22.6	42.6	49.4
FPN	✓		✓	38.9	60.3	41.9	22.7	42.6	49.8

4.3.2. Pyramid Scales

Table 3 shows the experimental results of the different settings of pyramid scales. It can be observed that single-scale GCDN (1st row) is inferior to multi-scale GCDN, indicating that multi-scale context representations are helpful to capture the visual features of objects with large size variations. Moreover, GCDN with larger pyramid scales is better at detecting large objects (AP_l increases from 1st row to 4th row).

Table 3. Impacts of different pyramid scales on MS-COCO 2017 *val*.

Method	Pyramid Scales	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
FPN	{1}	38.5	59.9	42.0	22.6	42.6	49.4
FPN	{1,2}	38.8	60.0	42.1	23.1	42.6	49.6
FPN	{1,2,3}	38.8	60.1	42.2	22.8	42.6	49.7
FPN	{1,2,3,6}	38.9	60.3	41.9	22.7	42.6	49.8

4.4. Lite Version

To reduce the computational cost, this work also presents a lite version of our method. For the lite version, the strategy to construct multi-scale context representations is changed. Specifically, for scale s , the global context X is divided into $s \times s$ subregions. For each subregion, we aggregate its contents by average pooling. By this way, the single-scale context representation $X^s \in \mathbb{R}^{C \times s \times s}$ is obtained. Both two version methods (i.e., lite and full) adopt the same pyramid scales of {1, 2, 3, 6}. Table 4 shows the experimental results. Compared with the FPN baseline, our lite version improves 1.1% AP without causing a big speed drop (runtime is evaluated on a single 2080 Ti GPU).

Table 4. Comparisons with lite version of our GCDN on MS-COCO 2017 *val*.

Method	Lite	Full	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l	Runtime FPS
FPN			37.4	58.1	40.8	21.4	40.8	48.5	15.3
FPN	✓		38.5	59.8	41.7	22.4	42.5	49.3	14.8
FPN		✓	38.9	60.3	41.9	22.7	42.6	49.8	14.2

5. Conclusions

This paper presents a novel Global Contextual Dependency Network (GCDN) framework, which captures global contextual dependencies to further enhance the local visual features of Regions of Interest (RoIs). The representation module and dependency module are designed to explore contextual information at different scales and generate contextual features, respectively. Comprehensive experiments validate that our approach brings consistent improvements for two-stage detectors and two modules complement each other. The key to capturing global contextual dependencies is to exactly model semantic and spatial relationships in the scene. Our next step is to efficiently capture global contextual dependencies.

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

Funding: This research was funded by the National Natural Science Foundation of China under Grant Nos. 61876069 and 62172185; Jilin Province Key Scientific and Technological Research and Development Project under Grant Nos. 20180201067GX and 20180201044GX; and Jilin Province Natural Science Foundation under Grant No. 20200201036JC.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.W.; Chen, J.; Liu, X.; Pietikäinen, M. Deep Learning for Generic Object Detection: A Survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
2. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
3. 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.

4. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R.S. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4898–4906.
5. Zhang, W.; Fu, C.; Xie, H.; Zhu, M.; Tie, M.; Chen, J. Global context aware RCNN for object detection. *Neural Comput. Appl.* **2021**, *33*, 11627–11639. [[CrossRef](#)]
6. Chen, Z.; Jin, X.; Zhao, B.; Wei, X.; Guo, Y. Hierarchical Context Embedding for Region-Based Object Detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 633–648.
7. Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; Wei, Y. Relation Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3588–3597.
8. Xu, H.; Jiang, C.; Liang, X.; Li, Z. Spatial-Aware Graph Relation Network for Large-Scale Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9298–9307.
9. He, J.; Deng, Z.; Zhou, L.; Wang, Y.; Qiao, Y. Adaptive Pyramid Context Network for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7519–7528.
10. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
11. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context; Lecture Notes in Computer Science. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8693, pp. 740–755.
12. Girshick, R.B.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
13. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
14. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Lecture Notes in Computer Science, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland, 2015; Volume 9351, pp. 234–241.
15. Gupta, D.K.; Arya, D.; Gavves, E. Rotation Equivariant Siamese Networks for Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 12362–12371.
16. Vedapant, N.; Wang, B. CRFace: Confidence Ranker for Model-Agnostic Face Detection Refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 1674–1684.
17. Redmon, J.; Divvala, S.K.; Girshick, R.B.; 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.
18. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; 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.
19. Girshick, R.B. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
20. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [[CrossRef](#)] [[PubMed](#)]
21. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; Murphy, K. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3296–3297.
22. Lin, T.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. 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. 936–944.
23. 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, 18–23 June 2018; pp. 6154–6162.
24. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
25. Galleguillos, C.; Belongie, S.J. Context based object categorization: A critical survey. *Comput. Vis. Image Underst.* **2010**, *114*, 712–722. [[CrossRef](#)]
26. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 213–229.
27. Li, J.; Wei, Y.; Liang, X.; Dong, J.; Xu, T.; Feng, J.; Yan, S. Attentive Contexts for Object Detection. *IEEE Trans. Multimed.* **2017**, *19*, 944–954. [[CrossRef](#)]
28. Wang, X.; Girshick, R.B.; 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, Y.; Wang, R.; Shan, S.; Chen, X. Structure Inference Net: Object Detection Using Scene-Level Context and Instance-Level Relationships. In Proceedings of the IEEE Conference on Computer vision And Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6985–6994.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *arXiv* **2017**, arXiv:1706.03762.
31. Bell, S.; Zitnick, C.L.; Bala, K.; Girshick, R.B. 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.
32. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
33. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.