symmetry

MDPI

Article

Object Detection by Attention-Guided Feature Fusion Network

Yuxuan Shi [1,†], Yue Fan [2,†], Siqi Xu [2,†], Yue Gao [3,†] and Ran Gao [2,*,†]

[1] School of Statistics and Data Science, Nankai University, No. 94 Weijin Road, Nankai District, Tianjin 300071, China; 2120210118@mail.nankai.edu.cn
[2] School of Journalism and Communication, Nankai University, No. 94 Weijin Road, Nankai District, Tianjin 300071, China; 1812759@mail.nankai.edu.cn (Y.F.); 2120201518@mail.nankai.edu.cn (S.X.)
[3] Zhou Enlai School of Government, Nankai University, No. 94 Weijin Road, Nankai District, Tianjin 300071, China; 1913309@mail.nankai.edu.cn
* Correspondence: gao.ran@nankai.edu.cn
† These authors contributed equally to this work.

Abstract: One of the most noticeable characteristics of security issues is the prevalence of "Security Asymmetry". The safety of production and even the lives of workers can be jeopardized if risk factors aren't detected in time. Today, object detection technology plays a vital role in actual operating conditions. For the sake of warning danger and ensuring the work security, we propose the Attention-guided Feature Fusion Network method and apply it to the Helmet Detection in this paper. AFFN method, which is capable of reliably detecting objects of a wider range of sizes, outperforms previous methods with an mAP value of 85.3% and achieves an excellent result in helmet detection with an mAP value of 62.4%. From objects of finite sizes to a wider range of sizes, the proposed method achieves "symmetry" in the sense of detection.

Keywords: object detection; feature fusion; attention

Citation: Shi, Y.; Fan, Y.; Xu, S.; Gao, Y.; Gao, R. Object Detection by Attention-Guided Feature Fusion Network. *Symmetry* **2022**, *14*, 887. https://doi.org/10.3390/sym14050887

Academic Editor: Dumitru Baleanu

Received: 31 March 2022
Accepted: 19 April 2022
Published: 26 April 2022

1. Introduction

Object detection is a fundamental and significant computer vision problem that seeks to recognize and locate all specified items in symmetric image processing. Convolutional Neural Network (CNN) and Deep Learning are heavily used in the current object detection frameworks [1]. Established approaches such as image processing, feature selection, and classification were used in prior studies. These approaches are classified into two types based on their detecting stages: one-stage methods and two-stage methods. The one-stage method is represented by the YOLO series network [2–4], while the two-stage method is represented by Faster-RCNN [5]. The main difference between the two methods is that the one-stage method has higher accuracy but lower speed.

In this case, YOLOv3 [4] achieves a very high level of speed. It employs a new backbone, DarkNet53, to extract image features. DarkNet53 is a deeper symmetric convolution neural network composed of several continuous residual structures [6]. This structure may reduce training difficulty while improving accuracy. YOLOv3's backbone is DarkNet53, which includes many residual blocks for extracting image features. YOLOv3 uses the Feature Pyramid Network (FPN) structure [7] to concatenate two different layers of features by upsampling a small feature map. In medium and large size objects, YOLOv3 performs significantly worse [8]. The use of local information is insufficient due to a lack of regional sampling, resulting in poor performance in some local area detection.

Accidents in the industrial sector are uncommon, but when they do occur, they frequently result in significant losses [9]. When it comes to preserving human health, few people perceive the impact of protective gear, and it's possible that it goes unnoticed. Workers' views of the necessity for safety helmets are inadequate, necessitating the development of more accurate safety helmet detection technology.

When two objects with different volumes are placed together, it is difficult to capture them at the same time. YOLOv3 is based on YOLOv1 and YOLOv2, but it improves detection accuracy, especially for small objects, while maintaining YOLO's speed advantage. When we detect employees wearing safety helmets on the construction site, the disparity in capture precision may result in omission, resulting in "Security Asymmetry".

We strive to avoid things being neglected owing to symmetric image capture accuracy in order to more correctly record objects of varied sizes in the same image. Although YOLOv3 does not offer a major improvement over YOLOv1 and YOLOv2, it can be used as a benchmark thanks to its high speed and enhanced accuracy. We are attracted to YOLOv3 high speed [4] and want to integrate it with our model to capture items of a larger range of sizes.

We consider that YOLOv3 has three different receptive field detectors for detecting big, medium, and small objects. The receptive field of these detectors is fixed. However, there are some objects whose scale is between big and medium or medium and small, in this case, YOLOv3 detectors have difficulty detecting these objects. We want a detector that can detect objects with scales ranging from large to small. To address this issue, we propose the Attention-guided Feature Fusion Network (AFFN) structure, which uses an attention module to guide feature fusion.

By combining our structure with YOLOv3, the goal of responding to any-scale object in a flexible manner has been attained. Then, we apply our work to helmet detection. Obviously, the accuracy of target detection has been greatly improved. Due to the use of the AFF module, our structure outperforms YOLOv3 in both subjective and objective evaluations.

To summarize, the contribution of this work is three-fold:

- **The method of feature aggravation is intended for extracting more representative features in object detection.**
- **An effective attention mechanism Channel-Spatial Attention Module (CSAM) guides feature fusion to improve detection accuracy.**
- **The proposed Attention-guided Feature Fusion Network (AFFN) method is capable of reliably detecting objects of a wider range of sizes, outperforms YOLOv3 in both subjective and objective evaluations. Furthermore, we apply our work on helmet detection and achieve excellent results, which contributed to operation security.**

2. Related Work

2.1. Object Detection in Convolution Neural Network

It has been shown that symmetric convolutional neural networks are effective in a variety of computer vision tasks. Symmetric convolutional neural networks are used in practically all object detection algorithms today. Two types of DCNN-based detectors exist: two-stage and one-stage. For the two-stage method, it first uses a convolutional network, such as VGGNet [10], to develop a region proposal with likely existing items, after which the area is calibrated and categorized to obtain the final result. Besides, a series of Region-based Convolutional Neural Networks (R-CNN) [11] approaches include Fast-RCNN [12] and Faster-RCNN [5], which are two-step algorithms with greater accuracy but slower speed.

Two-stage detectors introduce a Region-Proposal Network (RPN) to increase the efficiency of producing proposals and update the prediction positions of proposal regions, whereas one-stage detectors usually do not produce proposals. YOLO [3,4], for example, employs a symmetric convolutional neural network to make class and bounding box predictions directly. SSD, however, uses default anchors to adjust to different shapes of objects. YOLOv2 [3] has a built-in multi-scale training system and a new network structure called DarkNet-19. In addition, YOLOv3 [4] proposes a deeper residual network, Darknet-53, for better detection of small objects. Compared with one-stage and two-stage methods, proposal-free one-stage methods do not require pre-region classifiers.

2.2. Feature Aggravation

Feature aggregation is a common method for temporal applications like video comprehension [13], image super-resolution [14], and semantic segmentation [15]. The essence of the feature aggravation is to concatenate the symmetry features of several convolutions layers to improve the robustness of feature representation [16]. On picture super-resolution, a residual feature aggravation network [14] was used. Residual aggravation plays an important role in enhancing depict spatial detail. Feature aggravation was utilized in the symmetric video object detection task [13] to focus on features near the motion path to increase per-frame feature, allowing the video detector to capture moving objects faster and with greater accuracy.

2.3. Attention Mechanism

Attention mechanism is known as a network operator that creates features by fusing channel or spatial information in the immediate receiving domain [5,17]. Spatial-based attention, channel-based attention, and channel-spatial attention are examples of common attention processes. Channel attention learns varying weights on the channel dimension but constant weights on the plane dimensions, allowing it to focus on diverse parts of pictures. SENet [18] understands channels through learning. Based on the relationship, the importance of each typical channel is received. Effective features are louder in the process of learning. Moreover, in view of the learning results, unrelated features are held down. Spatial attention learns unique weights on the plane dimension, whereas channel attention learns the same weights. Cross domain attention is welcomed in recent works [19]. Many researchers are now being conducted using the channel-spatial attention mechanism. Convolutional Block Attention Module (CBAM) [20] extracts features by applying channel attention and spatial attention sequentially, whereas Bottleneck Attention Module (BAM) [21] applies spatial attention and channel attention on the features at the same time to obtain two distinct features, which are then concatenated. By explicitly modeling channel-wise and spatial feature interdependencies, the Channel-Spatial Attention Module (CSAM) [22] trains inter-channel and intra-channel feature responses. Gradient-weighted Class Activation Mapping (Grad-CAM) [23] is a convolutional neural network visualization that shows the properties learned by the convolutional neural network. Grad-CAM++ suggests an output gradient for weighting the pixel level at a specific position; this approach gives a measure of the relevance of each pixel in the feature map; and it provides a better explanation of object placement and the appearance of many objects instances on a single graph. Moreover, SPNet (Strip Pooling) [24] introduces a novel strip pooling paradigm that enables the backbone to catch long-distance dependencies successfully. In addition to spatial and channel dimensions, there is a temporal attention mechanism [25] that integrates time information to realize an attention summary for data containing time information. Recently, a new type of attention [26] that can achieve the same impact as spatial attention without additional parameters has arisen, supporting the study of attention-free mechanisms.

Attention modules are frequently used to enhance the performance of symmetric convolutional neural networks. We build Channel-Spatial Attention Module (CSAM) [22] on YOLOv3, which is a channel-spatial based attention mechanism that applies spatial and channel attention to the feature map concurrently while maintaining the feature map's original size. Single-channel attention and spatial attention are both restricted in their ability to identify objects. Channel attention always ignores spatial information, whereas spatial attention focuses on it. Therefore, these strategies have limits. Channel-spatial attention combines the advantages of channel and spatial attention. It may take into account both the channel and spatial information selection weights.

3. Method

3.1. Overall Network

In this section, we introduce the Attention-guided Feature Fusion Network (AFFN), which is based on YOLOv3 model. The specific implementation path is shown in Figure 1. The model consists of two main part: feature extraction part and object prediction part. First, the input image was fed into the feature extraction backbone to get variant scales of feature maps. DarkNet53 is used as the backbone. Then, different scales of extracted feature maps are aggravated together with three adjacent feature maps concatenated together and are input to the attention-guided feature fusion module to get channel and spatial attention information. The last three feature maps are processed by convolution blocks and then fused with the output of the attention module before convoluted to generate the corresponding prediction of objects. Taking the smallest group of generated feature maps as example, they first go into the convolution blocks, which are shown in Figure 2 and then are copied into two groups, the first group of feature maps are concatenated with the attention maps, which is obtained by incorporating the last three scale feature maps. After that, the concatenated feature maps will be used to generate the first level of object predictions by a 1×1 convolution and 3×3 convolution. The second group of feature maps are then upsampled and build together with the next to last size of the feature maps to get the second level size of features, following the above same steps.

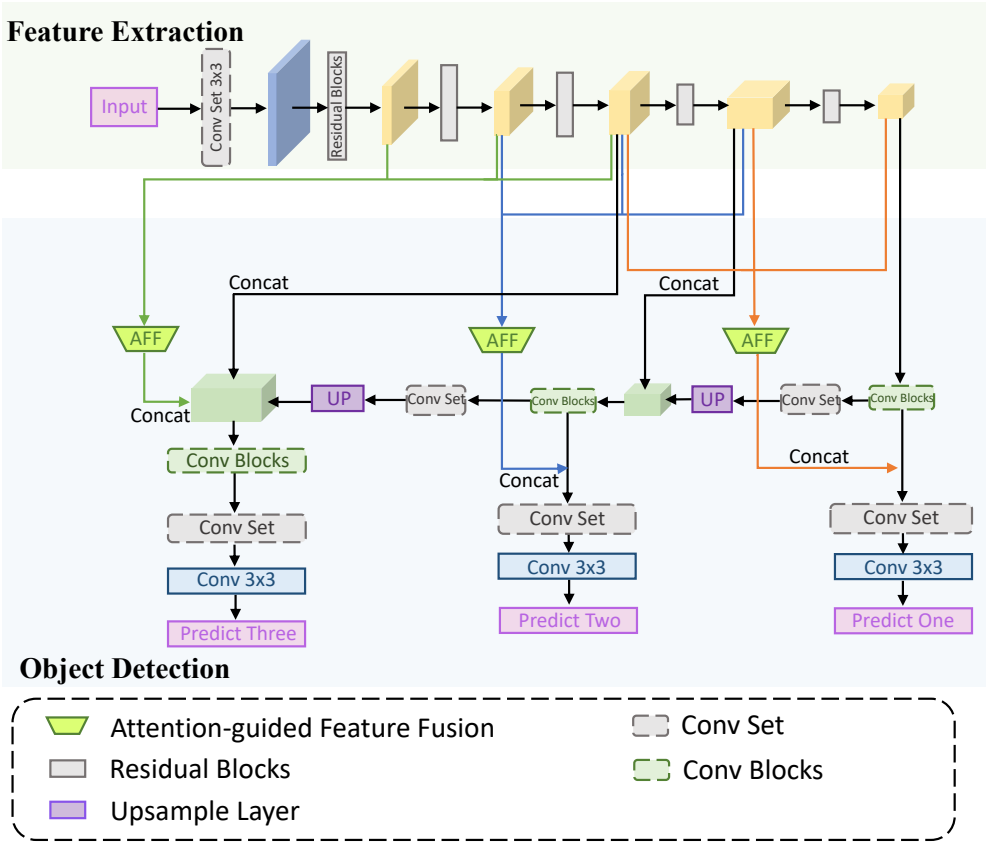

Figure 1. Demonstration of the Attention Feature Fusion Network (AFFN) architecture. The feature extraction uses the same DarkNet53 backbone in YOLOv3 with several Residual Blocks. Five different scales of feature maps are acquired after feature extraction and used in the object detection stage. The consecutive three groups of feature maps are collected and fused by Attention Feature Fusion (AFF) module to make full use of their channel and spatial information. Then the generated attention maps are taken in the corresponding object prediction output step to improve the object detection effect better. The upsample layer uses the nearest interpolation method to magnify the size of feature maps.

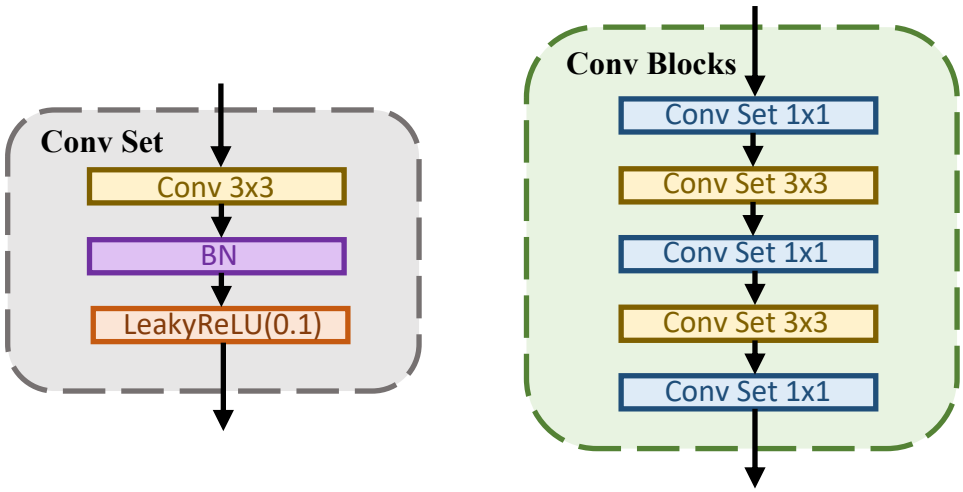

Figure 2. A further demonstration of "Conv Set" and "Conv Blocks" shown in Figure 1. The "Conv Set" consists of a convolution module, a batch normalization module and a LeakyReLU module with 0.1 as the negative gradient. The "Conv Set" has different size like 1×1 or 3×3 depending on the different kernel size of the inner convolution. Then, the "Conv Blocks" contains several "Conv Set" with different kernel size.

The core module in the proposed model is the attention-guided feature fusion structure, which could combine the adjacent three scales of feature maps and fuse them to get a wider range of object information. The concrete AFFN implementation structure is as following.

3.2. Attention-Guided Feature Fusion Module

Attention-guided feature fusion module can concatenate distinct layers' features which are then worsened by downsampling low receptive field ones, thus to concentrate high receptive field ones. Considering the collected features might be confused, we use a vision attention module to extract crucial information from aggravated features, which is a guider focus interest zone. The network operation logic of the AFF module is shown in Figure 3. We found that whereas single attention mechanisms such as channel attention or spatial attention focus on a single dimension feature, it is preferable to apply channel attention and spatial attention to the aggravating feature at the same time. To guide the aggravating feature, we use Channel-Spatial Attention Module (CSAM) [22]. It may reweight natural aspects and emphasize crucial characteristics of space and channels. Additionally, it may reweight natural characteristics and emphasize important ones in space and channels.

3.3. Implementation Detail

3.3.1. Network Structure

The whole network uses Darknet-53 as its backbone, and the entire Attention-guided Feature Fusion Module (AFFN) structure contains three AFF modules, each of which is added to the module before the detection results of each scale are produced. With reference to YOLOv3, after concatenating the results of the convolutional layer with the feature map of the previous scale, the results are concatenated and fed into the convolutional layer of the detection result output, which is used to extract feature maps of multiple scales.

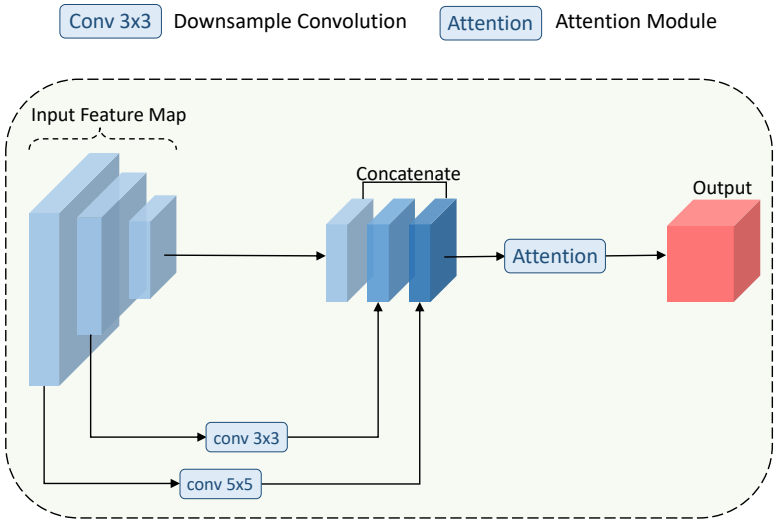

Figure 3. Illustration of the Attention-guided Feature Fusion (AFF) module. Three different sizes of feature maps are convoluted by a 5×5 convolution and 3×3 convolution separately and concatenated together. Then they are input into the channel and spatial attention module (CSAM) to fuse the feature scale information both in the channel and spatial level and get the output.

3.3.2. Loss Function

We utilize the iOU value to identify the detected target and a crucial value of 0.5 for the expected outcomes. And in order to identify the target function loss during training, we utilize the same Loss set as in YOLOv3 [4]. The predicted bounding boxes and center offset are constrained using Binary cross-entropy loss, the width and height of detection boxes are constrained using MSELoss [4], and the categorization of detection objects is constrained using Binary cross-entropy loss. Constraints on the confidence error are imposed by the cross-entropy loss. Otherwise, the ultimate loss function in network optimization is the total of the losses specified previously.

4. Experiments

4.1. Datasets

To train and assess the performance of our AFFN Module, we use Pascal VOC datasets [27]. Pascal VOC is a multifunction vision dataset that is used in the Pascal competition. It supports object identification, detection, and classification. 2007 trainval and 2012 trainval were assigned as train sets, and 2007 test was assigned as a test set. The test image's resolution is set at 544×544 pixels and the evaluation metric is mAP0.5.

4.2. Experiment Setting

We train our AFFN structure with pre-trained YOLOv3 weights on Pytorch framework [28]. DarkNet-53 [4] is the backbone we utilize. We train the network for a total of 50 epochs, with an 8 batch size. The Adam [29] optimizer is used to set an initial learning rate of 1×10^{-4} , a weight decay of 5×10^{-5}, a momentum of 0.9, and to alter the learning rate dynamically using cosine annealing learning rate.

4.3. Comparison Methods

We compare our AFFN Module with some representative object detection methods, including one-stage methods: YOLOv3 [4], SSD [30], DSSD [31], and two-stage methods: Faster-CNN [5], STDN [32], CoupleNet [33], and RFBNet [34].

4.4. Results on Pascal VOC Datasets

4.4.1. Objective Results

Table 1 shows the objective results. By comparison, our AFFN module is superior to existing representative methods, which include one-stage methods and two-stage methods with an mAP value of 85.3%. The YOLOv3 structure, which previously performs best, achieves an mAP value of 79.3%. When combined with our AFFN method, the performance of YOLOv3 is greatly improved with an mAP value of 84.3%. This confirms the validity of our proposed method. Compared with other networks, the AFFN method can achieve the goal of more accurate object detection.

Table 1. Detection Results on Different Methods.

Algorithm	BackBone	Test Images Size	Test Set	mAP@0.5
Faster-RCNN [5]	VGG16	1000 × 1000	VOC 2007	73.2
Faster-RCNN [6]	ResNet101	1000 × 1000	VOC 2007	76.4
SSD [30]	VGG16	300 × 300	VOC 2007	77.1
DSSD [31]	ResNet-101	513 × 513	VOC 2007	81.5
STDN [32]	DenseNet-169	513 × 513	VOC 2007	80.9
RFBNet [34]	VGG-16	512 × 512	VOC 2007	82.2
CoupleNet [33]	ResNet-101	1000 × 1000	VOC 2007	82.7
YOLOv3 [4]	DarkNet-53	544 × 544	VOC 2007	79.3
YOLOv3 (**Our**)	DarkNet-53	544 × 544	VOC 2007	**84.3**
AFFN (**Our**)	DarkNet-53	544 × 544	VOC 2007	**85.3**

4.4.2. Visual Comparison

Figure 4 shows the visual comparison results. We successfully use the AFFN structure to capture more features in the images and obtain objective results. Compared with the existing methods, the AFFN structure can respond to any-scale object in a flexible manner. YOLOv3 fails to generate clear structures. In contrast, our approach effectively suppresses such artifacts. For objects that can be detected in all these ten groups of images, our AFFN structure improves the accuracy by an average of 12% compared with the original method and increases the accuracy by 28% at the maximum. Moreover, we are successful in recognizing some items that YOLOv3 fails to detect in the images, such as the sofa in the third image and the cow in the tenth image.

4.5. Ablation Study

To examine the role of Feature Aggravation (FA) and Channel Spatial Attention Module (CSAM) in our AFFN structure, we utilize YOLOv3 as a baseline and test various combinations of these modules. Objective results are shown in Table 2. The candidates are Feature Aggravation (FA) and Channel-spatial Attention Module (CSAM).

Table 2. Ablation study of AFFN structure on VOC2007test. The resolution of test images is set to 544 × 544 pixels.

Backbone	Variant	Candidate		VOC2007test	
		FA	CSAM	mAP@0.5	mAP@0.75
DarkNet53	A			83.8	42.2
	B	√		84.8	44.1
	C		√	85.4	45
	D	√	√	85.8	45.7

Figure 4. VOC2007 Testset Results. (The first and the third row are YOLOv3 vs. results. The second and the fourth row are AFFN results.) We select ten groups of representative detection images. The differences in each comparison include cat, person, sofa, person, pottedplant, aeroplane, person, sofa, boat, and cow.

4.5.1. Effectiveness of Feature Aggravation and CSAM Attention Guidance

To determine the efficacy of feature aggravation, we disable the FA module and CSAM attention guidance module. Then we learn that target detection becomes less effective when we remove it. Therefore, it is necessary to add these two modules to make the detection effect better. As shown in Table 2, by comparing the results of "A" and "B", we discover that our AFFN structure with FA module obtains better performance. Similarly, by comparing "A" and "C", we can see that our structure with CSAM module performs better.

4.5.2. Influence of Different Number Feature Maps on AFFN

We add two, three, and four feature graphs with varying numbers to determine the influence on AFFN. We add two neighboring feature graphs into AFFN for two feature graphs. The three one is the same as the aforementioned instances. For the case of four, The final output keeps the three characteristics' input. As indicated in the Table 3, the optimum benefit is obtained by inputting the three neighboring characteristics into AFFN.

Table 3. Effect comparison of adding different numbers of feature maps to AFFN.

Number of Feature Maps	mAP@0.5	mAP@0.75
+2	83.93	43.08
+3	85.80	45.70
+4	83.88	43.05

4.6. Application on Helmet Detection

We apply our work to this task and see what happens. The aim of this training is to improve workplace safety by detecting people and hard hats. The datasets we use contain 5000 images with bounding box annotations in the Pascal VOC format for these three classes: Helmet, Person, and Head. The usability test is trained for a total of 50 epochs. The Adam [29] optimizer is used to specify a 1×10^{-4} learning rate and a 1×10^{-4} weight decay. Meanwhile, the freeze and unfreeze epochs are both 25. The evaluation metric is mAP@0.5.

4.6.1. Visual Comparison

Figure 5 shows the helmet detection results. We employ our AFFN structure to identify helmets, which are critical pieces of protective equipment in a variety of hazardous work situations. The present YOLOv3 detector is capable of locking by utilizing deep convolutional neural network features. We use the AFFN Module with YOLOv3, and it outperforms YOLOv3 in terms of helmet detection. Compared with YOLOv3, our AFFN structure can identify the helmets and heads that cannot be identified originally. For the helmets that can be correctly identified, our AFFN structure improves the average accuracy of 11%. The very small helmet like the one on the right of the fourth image is also successfully detected by our model.

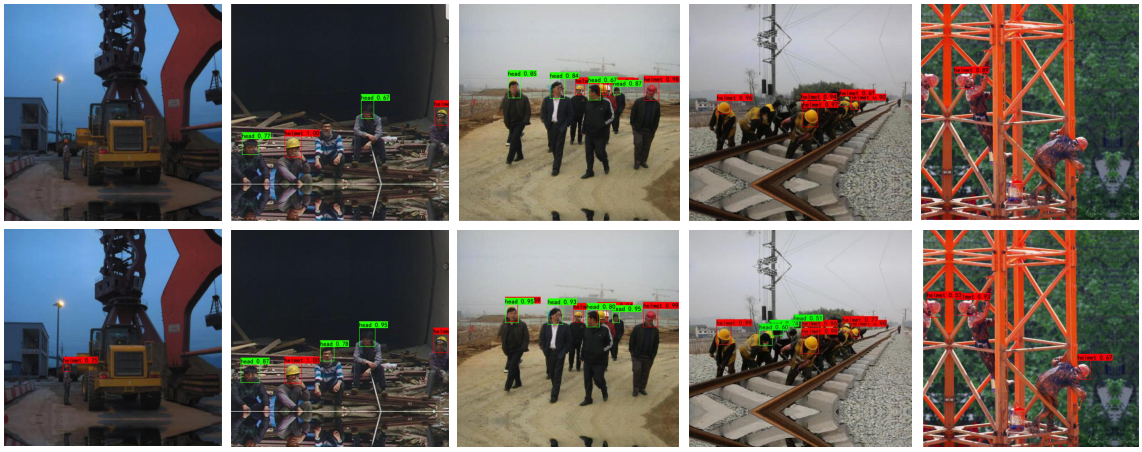

Figure 5. Helmet Detection Results. (The first row contains YOLOv3 results and the second row contains AFFN results).

4.6.2. Evaluation Results

The specific evaluation results are shown in Table 4. The result of YOLOv3 is 61.5, while our result is 62.4. Obviously, when we used AFFN Module with YOLOv3, the target detection became more effective.

Table 4. Evaluation Results on Helmet Detection.

Algorithm	Backbone	Test Images Size	Test Set	mAP@0.5
YOLOv3	DarkNet-53	544 × 544	Helmet test set	**61.5**
AFFN	DarkNet-53	544 × 544	Helmet test set	**62.4**

5. Conclusions

In this paper, we proposed an Attention-guided Feature Fusion Network and proved its effectiveness in dense object detection training. By applying our work to the helmet detection, efforts were also being made to enhance workplace security. In the YOLOv3 structure, the recognition impact of target features of three scales is obtained. For the diversity of objects, we integrated feature images of different sizes and applied a mixed

attention mechanism of channel and space to integrate feature information of multiple scales and realize object detection of more scales. Experiment results demonstrated that the AFFN Module is better than the previous methods, showing obvious superiority in image feature representation. Three nearby qualities of AFFN let it achieve an 85.3% mAP value, and the addition of the FA module and CSAM attention guidance module make it achieve the mAP value of 85.8%. An mAP value of 84.3% can be attained by combining our AFFN approach with YOLOv3 However, there are still certain concerns, such as insufficient generality of our research. Our future work aims at combining this approach with other networks to test its practicality further.

Author Contributions: Conceptualization, Y.S. and R.G.; Methodology, Y.S.; Software, Y.F.; Validation, Y.S. and R.G.; Formal analysis, Y.S.; Investigation, R.G.; Resources, S.X. and R.G.; Data curation, Y.F.; Writing—original draft preparation, S.X.; Writing—review and editing, S.X. and Y.G.; Supervision, R.G.; Visualization, S.X. and Y.G.; Project administration, Y.S., S.X. and R.G.; Funding acquisition, R.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds of the Central Universities and supported by the Supercomputing Center of Nankai University (NKSC).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in [PASCAL VOC 2007] at [http://host.robots.ox.ac.uk/pascal/VOC/voc2007/ (accessed on 6 November 2007), in [PASCAL VOC 2012] at [http://host.robots.ox.ac.uk/pascal/VOC/voc2012/ (accessed on 17 October 2012)] and in [Safety Helmet Detection] at [https://www.kaggle.com/datasets/andrewmvd/hard-hat-detection (accessed on 1 September 2021)].

Conflicts of Interest: The authors declare no conflict of interest. The funders have no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
2. 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.
3. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
4. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
5. 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*, 1–9. [CrossRef] [PubMed]
6. 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.
7. 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.
8. Hurtik, P.; Molek, V.; Hula, J.; Vajgl, M.; Vlasanek, P.; Nejezchleba, T. Poly-YOLO: Higher speed, more precise detection and instance segmentation for YOLOv3. *Neural Comput. Appl.* **2022**, *27*, 1–16. [CrossRef]
9. OMOYİ, C.; OMOTEHİNSE, A. A Factorial Analysis of Industrial Safety. *Int. J. Eng. Innov. Res.* **2022**, *4*, 33–43. [CrossRef]
10. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
11. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142–158. [CrossRef]
12. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
13. Zhu, X.; Wang, Y.; Dai, J.; Yuan, L.; Wei, Y. Flow-guided feature aggregation for video object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 408–417.
14. Liu, J.; Zhang, W.; Tang, Y.; Tang, J.; Wu, G. Residual feature aggregation network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2359–2368.

15. Li, H.; Xiong, P.; Fan, H.; Sun, J. Dfanet: Deep feature aggregation for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9522–9531.
16. Wu, Y.H.; Liu, Y.; Xu, J.; Bian, J.W.; Gu, Y.C.; Cheng, M.M. MobileSal: Extremely efficient RGB-D salient object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *12*, 1–11. [CrossRef] [PubMed]
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1–9. [CrossRef]
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
19. Zhang, Z.; Lin, Z.; Xu, J.; Jin, W.D.; Lu, S.P.; Fan, D.P. Bilateral attention network for RGB-D salient object detection. *IEEE Trans. Image Process.* **2021**, *30*, 1949–1961. [CrossRef] [PubMed]
20. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
21. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
22. Niu, B.; Wen, W.; Ren, W.; Zhang, X.; Yang, L.; Wang, S.; Zhang, K.; Cao, X.; Shen, H. Single image super-resolution via a holistic attention network. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 191–207.
23. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
24. Hou, Q.; Zhang, L.; Cheng, M.M.; Feng, J. Strip pooling: Rethinking spatial pooling for scene parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4003–4012.
25. Xu, G.; Xu, J.; Li, Z.; Wang, L.; Sun, X.; Cheng, M.M. Temporal modulation network for controllable space-time video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 6388–6397.
26. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In *International Conference on Machine Learning*; PMLR: New York, NY, USA, 2021; pp. 11863–11874.
27. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
28. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–12.
29. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
30. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
31. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
32. Zhou, P.; Ni, B.; Geng, C.; Hu, J.; Xu, Y. Scale-transferrable object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 528–537.
33. Zhu, Y.; Zhao, C.; Wang, J.; Zhao, X.; Wu, Y.; Lu, H. Couplenet: Coupling global structure with local parts for object detection. In Proceedings of the IEEE International Conference on COMPUTER Vision, Venice, Italy, 22–29 October 2017; pp. 4126–4134.
34. Liu, S.; Huang, D.; Wang, Y. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.