Helmet Detection Algorithm Based on Single Pixel Zoom

Yongheng Zhou, Lifen Jiang, Yan Liang, Chunmei Ma, Huazhi Sun, ShuaiBao Nie, Ying Zuo

College of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China
13132565021@163.com; wxxjlf@sina.com; mail_liangyan@163.com; memxhd@163.com; sunhuazhi@mail.tjnu.edu.cn; 670344856@qq.com; 1163558387@qq.com

Abstract. The automatic detection technology of helmet wearing is of great significance to the safety monitoring of construction site. It is difficult to detect a helmet when the object is small, overlapped, or unclear. Therefore, we propose SPZ-Det model to improve the performance of detect small targets in helmet detection task. We redesign the backbone network layers and proposes a Single Pixel Zoom (SPZ) module. SPZ-Det introduces the lower-level feature map, while for the same scale feature map, we select the higher-level feature map. Then the important elements of each feature map layer are amplified by SPZ module to solve the problem of small object feature fading. In the attention module of SPZ, the residual mechanism is introduced to weaken the influence of the attention. The proposed model improves the detection accuracy and keeps the detection speed. On the Safety-Helmet-Wearing dataset, the mAP of SPZ-Det is 27.9% higher than Efficientdet-d0 and gets an 8.8% increase compares with YOLOv3.

1. Introduction
Helmet is an important measure to protect head from injury or reduce the degree of head injury in construction sites. At present, production and business operation entities usually manually check the wearing of helmets by video monitoring or on-site inspection, which both require a large amount of human resources and cannot issue a timely warning to the workers who are not wearing helmets. In order to ensure the safety of workers in a construction site as much as possible, the automatic detection technology of wearing helmet has become a hot issue. One-stage detection algorithm YOLO[1] and SSD[2] has been widely used. However, YOLO[1] and SSD[2] are not effective in helmet detection, because there are the overlapped objects and small objects. In order to solve this problem, we select appropriate feature map layers, propose a Single Pixel Zoom module, and then combine with lightweight detection network Efficientdet[3] to build a new lightweight helmet wearing detection model. The concrete contributions of this paper are summarized as follows.

(1) We analyse the Efficientnet[4] feature layers, re-select the appropriate feature map layers, and introduce the lower-level feature map to supplement the small object feature which lost in the higher-level network, to improve the detection performance of small object.

(2) We propose a Single Pixel Zoom module(SPZ), which utilizes attention mechanism to strengthen the foreground elements in the feature map and weaken the background noise elements, so as to alleviate the phenomenon of feature disappearance in the process of network calculation. In addition, a residual structure is introduced into the attention module of SPZ to ensure the complete ability of characterization.
2. Related work
Helmet detection is the actual application field for object detection. At present, some scholars have conducted further research on automatic helmet recognition technology. Liu et al. [5] made use of the different colour distribution of the object pixels to identify the helmet. Zhang et al. [6] located the position of face by mapping skin colour, eye and mouth, then the face colour and shape features were extracted to judge whether to wear a helmet or not. These kinds of detection algorithm based on colour distribution relies heavily on the feature of colour difference of helmet, which was difficult to meet the detection environment with variable types of helmet.

This single feature dependence can be avoided in the end-to-end neural network detection algorithm. Actually, end-to-end object detection algorithm can be divided into two categories. One is the regression-based one-stage object detection algorithm, represented by YOLO[1], SSD[2] and RetinaNet[7], the other is the region-based two-stage object detection algorithm, such as Fast-RCNN[8] and Faster-RCNN[9]. Recently, Espinosa et al. [10] used Faster-RCNN[9] to detect the motorcycle helmet, but it is slow. And then Wu et al. [11] used one-stage detection algorithm YOLOv3[12] speed up to detect the helmet, but it adopts Densenet[13] as the backbone network needs to take up a lot of memory resources. We use the light one-stage detection network Efficientdet-d0[3] as the baseline to both achieve the detection speed and precision of the balance and reduce the consumption of computing resources. We propose a Single Pixel Zoom module to enhance the representation of object features with the thought of attention mechanism.

3. SPZ-Det Model
We have found the phenomenon of the small object is lost, and have developed a new Helmet detection models named SPZ-Det. We redesign the selection layers of the feature map in the backbone network to make the detection model suitable to the more complex environment. Meanwhile, a Single Pixel Zoom module (SPZ) is applied to improve the detection performance of the network. The network structure of SPZ-Det detection model is shown in Figure 1.

![Figure 1. Structure of SPZ-Det network model](image)

3.1. Feature Selection Layer
When we research the performance of the backbone network, we find that the feature representation ability of the same size feature layers is different. The lower-level feature map can supplement the details of small objects that are lost in higher-level feature map. Therefore, we redesign the feature map layers in the backbone network Efficientnet-b0[4]. The Efficientnet-b0[4] structure is shown in Figure 2 above.

Firstly, a lower-level feature map C2 is introduced to supplement the object detail features. For the C4 and C5 layers with the same feature scale, the high-semantic layers C4_2 and C5_2 are selected as the backbone feature layers to add the semantic features of the network. Finally, the backbone network feature map layers are C2, C3, C4_2, C5_2, and the top layer feature obtained by C5_2 down-sampling calculation and we call the final backbone network as Efficientnet-change.

![Figure 2. Efficientnet-b0 network](image)
3.2. Feature Zoom Layer
After feature extraction is carried out on feature map, it is necessary to carry out down-sampling or convolution calculation. In the calculation process, feature information of small objects may be replaced by other background features. Because the convolution kernel is difficult to capture enough information of small object. If the background features are more than the object features, which can lead to the information of object disappeared on the feature map and cause the error of object feature extraction.

We propose a Single Pixel Zoom (SPZ) module to reduce the missing probability of small objects. The SPZ module can enlarge the important elements and shrink the non-important elements to alleviate feature missing by calculating the contribution of each pixel in the feature map. In order to reduce the computational burden of the module, the GhostModule[14] is used to replace the original convolution operation. The SPZ module structure is shown in Figure 3.

![Figure 3. SPZ module structure. M is max-pooling; A is average-pooling; C is concatenate operation; S is the Sigmoid function; Ghost is the GhostModule[14]; Feature$_i$ is the formula 2.](file)

The SPZ module firstly performs a simple spatial attention calculation on the feature map to obtain the main area of the object. The value of attention can be calculated as formula 1.

$$score_i = S\left(\varphi(v(\text{max}(f_i), \text{mean}(f_i)))\right)$$

where $f_i$ is the vector-value in position $i$ over all channels, max is the Max-Pooling, mean is the Average-Pooling, $\left[\right]$ is the concatenation, $v$ is the convolution, then $\varphi$ and $S$ stand for $\text{Relu}$ and $\text{Sigmoid}$ function, respectively.

After the attention score is calculated by formula 1, the score is calculated by element-wise multiplication with the feature map. However, since the attention score has been compressed between $(0,1)$ by $\text{Sigmoid}$ function, the representational ability of the feature will be weakened. Thus, we utilize the residual mechanism to retain the original ability of the feature map.

Pixel-level feature zoom is performed on the feature after the enhanced attention feature map is obtained. The contribution value of each pixel to the feature map is calculated, and then the primary and secondary elements are scaled according to the contribution value. The specific calculation process can be formulated as formula 2.

$$Feature_i = \begin{cases} G(n_iC_i) + C_i & \text{if } S(C_i) \geq \frac{1}{H \times W} \\ G((1 - n_i)C_i) + C_i & \text{if } S(C_i) < \frac{1}{H \times W} \end{cases}$$

where $C_i$ is the feature of $i$ channel, $G$ is GhostModule[14], $n_i$ is the zoom value and $S$ is the Softmax that calculates the score for each channel. The score means the contribution of each pixel position to the channel feature, it will be compared with the average contribution score $\frac{1}{H \times W}$, where $H$ and $W$ are the height and width of the feature map. If the contribution score is not less than the average contribution value, the zoom value is set to $n_i$, else it is set to $(1 - n_i)$. And $n_i$ is a parameter automatically learned by the network, initialized to 0.9. Then the element-wise multiplication will be performed to get the zoom feature map. Finally, $C_i$ is added to obtain the final feature.

3.3. Loss function
In the final stages of network inference, the classification and location of the object are predicted by the detection head network. For the classification subtask, we use the Focal loss[7] strategy to calculate the classification loss, as shown in formula 3.

$$\text{Cls. loss}_i = -\alpha_i[y_i \times \log(cls_i) + (1 - y_i) \times \log(1 - cls_i)]$$

3
where $y_i$ is the real label of $i$ category, $cls_i$ is the prediction score of $i$ category and $\alpha_i$ is a balance factor.

For the location regression subtask, the regression value of anchor should be predicted to narrow the difference between anchor and the ground-truth box. The regression label $(dx, dy, dw, dh)$ is replaced by the relative position offset between the ground-truth box $(tx, ty, tw, th)$ and the anchor box $(ax, ay, aw, ah)$. The specific conversion rules are shown in formula 4.

$$
dx = \frac{tx - ax}{aw}; \quad dy = \frac{ty - ay}{ah}; \quad dw = \log \frac{tw}{aw}; \quad dh = \log \frac{th}{ah}
$$

(4)

The final regression loss uses smooth L1 function strategy, and the actual implementation is shown in formula 5.

$$
Reg_loss = \begin{cases} 
0.5 \cdot |gt - reg|^2 & |gt - reg| < 1 \\
|gt - reg| - 0.5 & \text{others}
\end{cases}
$$

(5)

where $gt$ is the regression offset after conversion, $reg$ is the offset predicted by the regression subnetwork. The total loss function of the network is the sum of classification loss and regression loss.

4. Experiments

4.1. Dataset

The open Safety-Helmet-Wearing-Dataset provided by Pactera is used in our experiments. There are 7582 images in the data set, which contains 9044 objects wearing safety helmet bounding box (positive) and 111514 objects not wearing safety helmet of bounding box (negative).

4.2. Experiment Settings

We use two 12GB Nvidia GeForce GTX 1080Ti GPUs to train the network. And backbone network uses Efficientnet-b0[4] pre-trained parameters to speed up the convergence. We choose the adjustment strategy associated with learning rate and every training result. In other words, the learning rate will decrease when the training loss does not change after the patience time continuous training, and we set patience to 3. Then the ratio of anchor is $[0.5, 1, 2]$, the scales set to $[1, 2^\frac{1}{4}, 2^\frac{1}{2}]$ and the epoch is 150.

4.3. Result analysis

Our proposed model SPZ-Det improves the performance of Efficientdet[3] by reselecting feature map layers and introducing SPZ module. We make an ablation experiment to compare performance between the Efficientdet-d0[3], the Efficientdet-change network and our SPZ-Det. The experimental results are shown in Table 1.

| Network          | AP (hat) | AP (person) | mAP | fps | params |
|------------------|----------|-------------|-----|-----|--------|
| Efficientdet-d0  | 79.6%    | 24.9%       | 52.3% | 22  | 15.9M  |
| Efficientdet-change | 93.9%   | 61.9%       | 77.9% | 20  | 15.9M  |
| SPZ-Det          | 94.6%    | 65.8%       | 80.2% | 20  | 16.1M  |

the Efficientdet-change is the model which uses Efficientnet-change as backbone network of Efficientdet-d0. As we can see from Table 1, the AP of Person class improves significantly after the introduction of the Efficientnet-change. Because people class are mostly small objects in this dataset, the introduction of the bottom feature layer will enrich features of small object greatly. Furthermore, after the Single Pixel Zoom module is utilized in the network, the accuracy is further improved 2%.

In order to prove the effectiveness of the Single Pixel Zoom module, we add this module to the YOLOv3[12] network and conduct experiments on the same dataset. The experimental results are shown in Table 2.
Table 2: YOLOv3 contrast experiment

| Network     | AP (hat) | AP (person) | mAP  |
|-------------|----------|-------------|------|
| YOLOv3      | 86.3%    | 56.4%       | 71.4%|
| YOLOv3+SPZ  | 91.5%    | 55.5%       | 73.5%|

The comparative analysis of the experimental results shows that Single Pixel Zoom module can be embedded into other detection models, improving the detection performance of the model.

![Figure 4](image)

Figure 4. Actual detection example

In Figure 4, the red box represents a head worn helmet and the blue box represents a head not worn helmet. In figure 4(a), a head without a helmet is detected and classified correctly. In figure 4(b) and 4(d), the occluded objects far from the camera all can be identified correctly. Based on the analysis of the detection results, it can be concluded that the SPZ-Det model performs well to detect small objects and occluded objects, which can meet the detection requirements of the actual scenes with complex personnel and changeable locations.

5. Conclusion

In our paper, a Single Pixel Zoom detection model (SPZ-Det) is proposed, which is designed for complex occlusion and small object detection. The model introduces the lower-level feature map with rich details into the network and ensures the effectiveness of the network in detecting small object. Single Pixel Zoom module (SPZ) enhances the important information in the feature, which ensures that the important feature information will not be ignored or replaced by other noise features when the network is inferencing. Based on the selection of feature map layers and the introduction of SPZ module, the loss of feature information in network transmission is solved. The effectiveness of SPZ module is proved by comparison experiments. Our model improves the accuracy of object detection based on ensuring the detection speed. The detection accuracy in wearing the helmet reached 94%, which can meet the requirements of the safety helmet monitoring under the basic operating environment.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant No.61702370, Tianjin Natural Science Foundation under Grant No.18JCYBJC85900, 18JCQNJC70200, Tianjin Science and Technology Project under Grant No.17ZLZXZF00530, 131 Three-Level Personnel of Tianjin Normal University under Grant No.043/135305QS20.

References

[1] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
[2] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.
[3] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10781-10790.
[4] Tan M, Le Q V. Efficientnet: Rethinking model scaling for convolutional neural networks[J].
[5] Yunbo Liu, Huang H. Research on monitoring of workers’ helmet wearing at the construction site[J]. Electronic Science and Technology, 2015, 28(4): 69-72.

[6] Zhang G, Lv L, Li D, et al. The Method for Recognizing Recognition Helmet Based On Color and Shape[C]//2017 5th International Conference on Machinery, Materials and Computing Technology (ICMMCT 2017). Atlantis Press, 2017.

[7] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[8] Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(1): 142-158.

[9] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

[10] Espinosa J E, Velastin S A, Branch J W. Motorcycle detection and classification in urban Scenarios using a model based on Faster R-CNN[J]. 2018.

[11] Wu F, Jin G, Gao M, et al. Helmet Detection Based On Improved YOLO V3 Deep Model[C]//2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC). IEEE, 2019: 363-368.

[12] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[13] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[14] Han K, Wang Y, Tian Q, et al. GhostNet: More features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1580-1589.