Non-Standard Clothing Detection in Electricity Scenes Based on Adaptive Training Samples Selection Neural Network

Haipeng Chen\textsuperscript{1,3}, Wang Luo\textsuperscript{1,2*}, Jinwei Mao\textsuperscript{1,2}, Fuhon Min\textsuperscript{3}

\textsuperscript{1}NARI Group Corporation (State Grid Electric Power Research Institute), Nanjing, Jiangsu, China
\textsuperscript{2}Nanjing NARI Information and Communication Technology, Co., Ltd, Nanjing, Jiangsu, China
\textsuperscript{3}Nanjing Normal University, School of Electric Engineering and Automation, Nanjing, Jiangsu, China

\textsuperscript{*}email: luowang@sgepri.sgcc.com.cn

Abstract. Standard clothing in the electricity scene is so important that it can validly prevent workers from injuries. Strengthen non-standard clothing detection can help people correct the bad habits of dressing. However, there are some challenges existing in standard clothing in electricity scenes. There are few numbers in some kinds of clothing (e.g. short sleeves or not wearing helmet) in electricity scenes. The colors of some personnel clothing are similar to those of the background of workshop, such as white helmets and white frames. To handle these issues, in this paper, a novelty clothing detection method in electricity scenes is proposed based on a policy of adaptive training samples selection. The number of clothing can be expanded by rectifying mosaic augmentation, and the information loss of the top features can be compensated by residual feature augmentations. Experimental results show that the model can automatically and accurately identify non-standard clothing from complex electricity scenes, and has better detection performance (mAP=0.433) compared with the other model (mAP=0.419).

1. Introduction

Electricity scenes belong to high-risk workplaces. Especially in the construction phase, some enterprises leave steel pipes, cables everywhere to save time, making a threat to personal safety. Standard clothing can reduce personnel injuries, but workers often have little awareness of safety protection. According to the statistics, in recent years, the accidents caused by incorrect wearing of safety protective equipment in the operation process account for more than 50\% of accidents in the power industry [1]. At present, the electricity scene generally adopts manual safety monitoring, but people are susceptible to external factors and may not be able to concentrate resulting in outdated global monitoring, which will lead to safety accidents [2]. Therefore, it is urgent to strengthen the inspection of non-standard clothing.

There have been a lot of research on non-standard clothing at home and abroad. Literature [3] proposes to extract the shape and color feature vector of the clothing of workers, to apply the Monte Carlo method to random sample point in the field of training samples, and to build a loss function considering neighborhood sensitivity. Gauss Newton iteration method is employed to solve the weights from hidden layer to output layer. However, the detection performance on clothing of squatting personnel is poor. Literature [4] proposes to split the images of person into three cells
(helmet, coat and trousers), and features of the cells are extracted by HOG and HOC to train the linear SVM classifier.

Through the introduction of deep learning, the performance of non-standard clothing detection algorithm has been greatly improved. Literature [5] proposes to introduce self-attention mechanism into the YOLOV3 [6] model to mine the hidden features and strengthen dependency relations among classes. It helps to detect the similar clothing. Literature [7] combines Faster R-CNN [8] with the algorithm of Deep Feature Flow [9] to detect the key frame images of the video, so as to realize the detection of safety helmets in images and videos.

To expand number of samples, SSD [10] proposes several data augmentation which can be categorize into two aspects (photometric distortions and Geometric Distortions). Photometric distortions include random brightness, random contrast, hue, saturation and random lighting noise. Geometric distortions include random expand, random crop and random mirror. But there are some limits on expanding number of sample, it concentrates on single sample data augmentation. To strengthen context feature, Feature Pyramid Network (FPN) [11] builds a feature pyramid upon the inherent feature hierarchy in ConvNet by propagating the semantically strong features from high levels into features at lower levels.

In this paper, we propose an Adaptive Training Sample Selection V2 (ATSS_V2) model, which is based on Adaptive Training Sample Selection (ATSS) and improved by rectified mosaic data augmentation [12] and residual feature augmentation [13]. Rectified mosaic data augmentation compared with SSD (which is only single sample data augmentation) includes not only single sample data augmentation but also multi sample data augmentation, and it deletes the noise boxes. Residual feature augmentation recovers information at the highest level compared with FPN (which ignores the information loss of features at the highest level so that it extracts insufficient context feature). Experiments show ATSS_V2 is more robust and accurate compare with ATSS model in electricity scenes.

2. Models

2.1. Based model ATSS

ATSS is famous for an adaptive training sample selection strategy to select anchors. The structure of ATSS can be seen in Figure 1. For each Ground Truth (GT) of input images, L2 distance between the center point of GT and the central point of the preset anchor is calculated at the feature levels of P2, P3, P4 and P5 separately, and the 9 closest preset anchors in each feature level are selected as candidates. At each feature level, multi Interaction over Union (IOUs) between 9 preset anchors and the GT are calculated respectively. As shown in Figure 2, IOUs corresponding to four feature level are b1, b2, b3, b4 respectively. Mb is the average of IOUs, and Vb is the standard deviation of IOUs. The final IOU threshold is Mb + Vb. When the IOU between the candidate and the GT is greater than the final IOU threshold and it meets the condition of distance from the center point of GT, the candidate will be regarded as positive training sample of the GT, or they will be negative training sample. If the candidate corresponds to multiple GTs and the IOU between the candidate and one of the GT is the highest, the candidate is set to be the positive training sample of the GT, and the candidate is a negative training sample for the remaining GTs.
Fig. 1 Structure of ATSS

As shown in Figure 2 (a), when Mb is too small, most of the candidate of GT are of low quality. When Vb is too small, multiple feature levels are suitable for detecting this object, so the IOU threshold during training should be set lower. As shown in Figure 2 (b), when Mb is big, most of the candidate have better detection performance of the object. When Vb is big, it indicates only P3 feature level is suitable for detecting the object. By setting eventual IOU thresholds for each GT, the model can choose the appropriate positive training samples from the appropriate level of the feature levels, and the rest are the negative training samples.

Fig. 2 Settings of IOU

2.2. Methodology

Although ATSS succeeds in searching for correct training samples, it is not very valid in feature fusion. Because it adopt FPN in feature fusion, in other words, it ignores the information loss of features at the highest level. P5 feature level in FPN suffers from the information loss due to the reduced feature channels and only supports single scale context information. In order to strengthen features at P5 feature and enrich samples, we propose ATSS_V2.

2.2.1. Model structure of ATSS_V2. The model structure is shown in Figure 3, including 4 parts: feature extraction, residual feature augmentation, feature fusion, and head. First, we use improved mosaic as data augmentation to expand numbers of input, and apply Resnet-50 backbone to extract the features at feature levels, such as C2, C3, C4 and C5. Second, we apply residual feature augmentation module at C5 feature level to get recovered features level C6. Third, combine the features at C6 feature level and the features at C5 feature level with residual connection to get the final prediction features at P5 feature level. Finally, classification and regression tasks are carried out on features at feature levels (P2, P3, P4, and P5).
2.2.2. Data augmentation. It can be seen from Figure 4 that there are four original images. We let (a) as $x_1$, (b) as $x_2$, (c) as $x_3$, (d) as $x_4$, generated image as $\bar{x}$, single sample data augmentation function as $f(x)$, and $f(x)$ can be obtained by zooming original input images, changing the color gamut of original input images, cropping original input images (cut the upper left corner of the generated image in the upper left direction, the lower left corner of the generated image in the lower left direction and so on).

The mosaic data augmentation function is

$$\bar{x} = f(x_1) + f(x_2) + f(x_3) + f(x_4)$$  \hspace{1cm} (1)

![Fig. 3 Structure of ATSS_V2](image)

![Fig. 4 Original images](image)
To be specific, we preset the size of the generated image \((H*W)\) in Figure 5 (a), and then divide the generated image into four blank modules according to the two cutting lines \((\text{cut}_x\text{ and cut}_y)\) of the generated image, and then fill four blank modules with four processed images \((f(x_1), f(x_2), f(x_3), f(x_4))\) respectively. Then place the processed images in four different directions of upper left, lower left, lower right and upper right respectively to get the generated image (Figure 5b). To save the time of annotating on generated images, we read the annotations of four original images, and generate the annotations of generated images by means of scaling in the same proportion in which the images are scaled, keeping the same offsets of box as those of original images. This is the principal of mosaic data augmentation.

But there is a drawback in mosaic data augmentation. It remains boxes which aren’t beyond the transformed image \((f(x_2))\), but maybe parts of original boxes \((x_2)\). It can be seen from people in the right corner of \(f(x_2)\) in Figure 5(b) that the label of person is nearly overlapped with the label of helmet resulting in confused classification. To handle this, we delete the boxes which are parts of original boxes in the generated images and get final generated images as Figure 5(c) shows. To learn more information, our datasets after improved mosaic data augmentation consist of generated images and original images.

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{Class} & \text{Before}_\text{}\text{Augment} & \text{Before}_\text{}\text{Augment} & \text{After}_\text{}\text{Augment} & \text{After}_\text{}\text{Augment} \\
\text{Picture Nums} & \text{XML Nums} & \text{Picture Nums} & \text{XML Nums} \\
\hline
\text{helmet} & 947 & 2529 & 2411 & 10570 \\
\text{long_sleeve} & 1113 & 2125 & 2591 & 8135 \\
\text{no_helmet} & 94 & 736 & 1112 & 3337 \\
\text{person} & 1244 & 3227 & 2676 & 9123 \\
\text{short_sleeve} & 99 & 113 & 391 & 469 \\
\text{trousers} & 1054 & 2405 & 2422 & 7480 \\
\text{total} & 3951 & 11115 & 11603 & 39114 \\
\hline
\end{array}
\]

\textbf{Fig. 5} Data augmentation

Results after improved mosaic data augmentation (Mosaic’) can be seen in Table 1. Experiments show that the number of some kind of clothing (no_helmet, short_sleeve) XML increased amazingly from 736 to 3337 and from 113 to 469. It enriches the sample number in some extent.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
\text{Class} & \text{Before}_\text{}\text{Augment} & \text{Before}_\text{}\text{Augment} & \text{After}_\text{}\text{Augment} \\
\text{Picture Nums} & \text{XML Nums} & \text{Picture Nums} & \text{XML Nums} \\
\hline
\end{tabular}
\end{table}

\textit{2.2.3. Information recovery of highest feature level.} In order to make up for the information loss, this paper adopts the residual feature augmentation module, whose structure is shown in Fig. 6. It consist of several components. Assuming that the size of C5 is S, after adaptive pooling(rate r1, r2, r3 is 0.1, 0.2 and 0.3 respectively) and 1*1 conv operation, we get three kinds of features whose dim N is 256. We concatenate these three kinds of features and generated the feature whose dim becomes NC. Generated feature is input to generated weight module. In the generated weight module, by applying 1*1 conv, feature’s dim is reduced from NC to C. And after 3*3 conv, feature’s size become N*H*W.
We use sigmoid fuction to generate a weights list whose shape is N*(1*H*W) and upsample the features in generated features module to a scale of S via bilinear interpolation. Finally, we combine the upsampled features with different weights generated in the generated weight module to get the new feature level C6.

![Diagram of residual feature augmentation module]

**Fig. 6** Residual feature augmentation module

### 3. Experiments

#### 3.1. Experimental environment

The computer is configured with Ubuntu18.04 operating system, CPU Genuine Intel, GPU Quadro GV100, memory 128G, Python 3.8 and Pytorch 1.5 framework. The Quadro GV100 has a volty architecture with 5120 CUDA cores, a single precision floating-point performance GPU BOOST which is up to 14.8 Teraflops and a double precision floating-point performance which is up to 7.4Teraflops.

#### 3.2. Experimental evaluation standard

In the detection task of non-standard clothing, Average Precision (AP) is adopted in this paper. AP can reflect the detection precision of the model.

#### 3.3. Ablation study

We run a number of ablations to analyze ATSS V2. We show the ablation experiments on detailed designs of each component. Results are shown in Table 2.3 and discussed in detail next.

| Method          | mAP  | AP50  | AP75  |
|-----------------|------|-------|-------|
| ATSS            | 0.419| 0.819 | 0.362 |
| ATSS+C6         | 0.420| 0.824 | 0.376 |
| ATSS+Mosaic'    | 0.420| 0.825 | 0.368 |
| ATSS+C6+Mosaic' | 0.433| 0.836 | 0.378 |
Tab. 3 Map when applied different methods

| Method              | mAP_s | mAP_m | mAP_l |
|---------------------|-------|-------|-------|
| ATSS                | 0.290 | 0.278 | 0.473 |
| ATSS+C6             | 0.164 | 0.287 | 0.484 |
| ATSS+Mosaic’        | 0.097 | 0.304 | 0.483 |
| ATSS+C6+Mosaic’     | 0.274 | 0.331 | 0.481 |

Fig. 7 Confusion matrix analysis
It can be seen from Table 2 that when we adopt C6 feature or Mosaic', AP50 and AP75 increase obviously, but mAP increases a little. When we adopt C6 feature and Mosaic' at the same time, AP increases by 1.4 point, and AP50 and AP75 increase obviously. This can be understood by Table 3. When we adopt C6 feature or Mosaic', the performance of small objects drops sharply, and performance of medium objects and large objects increases. In other words, C6 feature or mosaic' helps detect medium objects and large objects but fails to detect small objects. When we adopt C6 feature and mosaic' at the same time, we can retain the performance of small objects and the performance of medium objects and large objects continue to increase.

Fig. 7 is the analysis of confusion matrix. Its x-coordinate is the prediction label, and y-coordinate is ground-truth label. Diagonal lines represent the number of correctly identified categories, and non-diagonal lines represent the number of incorrectly identified categories. Figure 7(a) is ATSS+C6+Mosaic', Figure 7(b) is ATSS+C6, Figure 7(c) is ATSS+Mosaic', Figure 7(d) is ATSS. It can be inferred that it predicts wearings (helmet, no_helmet, person, trouser) more accurate than other models. Figure 8 is our visible experimental result, it seems our proposed model can detect the wearing well.

4. Conclusion
In this paper, in order to accurately identify personnel clothing from the electricity scenes, we proposed an neural network model namely ATSS_V2. We conduct improved mosaic data augmentation, residual feature augmentation module to expand numbers of samples, improve feature expressions respectively. Experiments show that improved data augmentation and residual feature augmentation can help to detect personnel clothing.

Acknowledgments
This work was funded by the State Grid Corporation Science and Technology Project " Research on FPGA intelligent Computing technology for power image and video understanding " (Grant No. 5700-202119266A-0-0-00).

References
[1] Yu J and Zhang Y 2018 Proc. Int. Conf. on National Steel Structure Engineering Technology (Hangzhou) vol 47 pp 441-444
[2] Chang Z, Peng Q and Chen Y 2020 Safety supervision method based on machine learning and image recognition in electrical field China Electric Power 53 pp 155-160
[3] Li C, Hu J, Zhou J and Kong W 2019 Image recognition algorithm for transformer operators based on S-RBFNN Guangdong Electric Power 32 pp 125-129
[4] Hu J, Zhou J, Lin X, Jiang H and Li C 2018 Research on clothing analysis method of substation operator based on S-HOG+C Operator Mechanical and Electrical Engineering Technology
[5] Mo B, Wu K, 2020 Research on Intelligent detection technology of illegal wear in electric power operation by introducing self-attention Computer and Modernization 294 pp 115-121

[6] Redmon J and Farhadi A 2018 YOLOv3:An Incremental Improvement arXiv (preprint arXiv:1804.02767)

[7] Deng K and Zhou Z 2020 Implementation and analysis of helmet wearing detection based on deep learning Computer Era 7 pp 12-21

[8] Ren S, He K, Girshick R and Sun J 2015 Neural Information Processing Systems (Montréal)

[9] Zhu X, Xiong Y, Dai J, Yuan L and Wei Y 2017 IEEE Conference on Computer Vision and Pattern Recognition (Boston)

[10] Liu X, Zhang B, Fu Y and Zhu J 2020 Deep learning based dressing standardization test for workers in polluted sites China’s Science and Technology of Production Safety 7 pp 169-175

[11] Lin T, Dollár P, Girshick R, He K, Hariharan B and Belongie S 2017 Feature Pyramid Networks for Object Detection arXiv (preprint arXiv:1612.03144)

[12] Bochkovskiy A, Wang C and Liao H 2020 YOLOv4: Optimal Speed and Accuracy of Object Detection arXiv (preprint arXiv:2004.10934)

[13] Guo C, Fan B, Zhang Q, Xiang S and Pan C 2020 AugFPN: Improving Multi-scale Feature Learning for Object Detection arXiv (preprint arXiv:1912.05384)