Person detection and identification surrounding gas facilities based on Improved-Yolov4-tiny

Zikang Wang\textsuperscript{1a}, Yufeng Zhuang\textsuperscript{1b}, Jiale Li\textsuperscript{1c}, Xingde Wang\textsuperscript{2d}

\textsuperscript{1}School of Modern Post (School of Automation), Beijing University of Posts and Telecommunications, Beijing, China
\textsuperscript{2}School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
\textsuperscript{a}wangzikang@bupt.edu.cn, \textsuperscript{b}zhuangyf@bupt.edu.cn, \textsuperscript{c}jil19971013@163.com, \textsuperscript{d}wxd_sander@bupt.edu.cn

Abstract—The safety of the gas facilities is an important guarantee for the production and life of the society, and the monitoring of people entering the area is an important means to guarantee the safety of the facility. In order to solve the problem of non-contact monitoring of people entering the gas facilities, an Improved-Yolov4-tiny method is proposed in this paper. This method can be used to monitor people entering the area in real time and determine whether the type of person is a worker or non-worker based on characteristics such as clothing. Among them, the model increases a large-scale feature layer to improve the recognition and detection of small targets. The introduction of attention mechanism for multi-scale feature fusion increases the effectiveness of fine-grained detection of person. The Improved-Yolov4-tiny detection algorithm not only meets the requirements of real-time monitoring, but also achieves better results at different resolutions. Compared with the traditional Yolov4-tiny detection algorithm, the Improved-Yolov4-tiny detection algorithm improved the mAP metric by 3.1%.

1. INTRODUCTION
With the rapid development of society and economy, gas has become one of the important energy sources in people's daily life. Gas facilities are important tools and carriers for gas transportation, and cities have established huge gas facilities to ensure the safety of gas transportation is of great importance to the whole society [1]. Gas facilities are an important part of the energy system. In recent years, vandalism of gas facilities has occurred from time to time. When non-workers enter the area, there are safety risks to the gas facilities and they need to be monitored for conditions affecting the safe operation of the facilities in the area where they are operating to prevent damage to the facilities and theft.

With the development of technology, the number of video surveillance in the area of gas facilities has been growing rapidly, and the way of monitoring by manual guards can no longer meet the growing requirements. A complete and efficient intelligent monitoring system can effectively guarantee the safety of gas transportation. Person detection is an important part of the intelligent monitoring system and provides support for the intelligent monitoring system. Detecting persons entering the facility area and determining the type of identity of persons entering the facility area based on the characteristics of their clothes is important for intelligent monitoring and securing the safety of gas transportation.
Currently, the methods of person detection can be divided into two main categories. The first category is the artificial feature-based person detection method, which uses Haar [2] features, HOG [3] features, HOG-LBP [4] features, and SIFT [5] features as person features, and then trains a classifier with the extracted features for recognition detection. For example, in order to solve the problem of detecting unsafe behaviors of subway construction personnel, Xie Yi et al., divided the problem into three parts, in which person detection was performed using a combination of support vector machine (SVM) and histogram of directional gradients (HOG) features [6]. The second category is based on deep learning methods. Extracting features for detection using target detection models. Yang Shichao used Faster-Rcnn network for person detection in underground mines, capturing personnel images by camera, inputting the images into the neural network for feature extraction, classification and coordinate regression, and finally obtaining an accurate person detection map [7]. Ivasic-Kos et al., used the improved YOLOv3 algorithm to detect pedestrians in thermal imaging and achieved better results with guaranteed speed [8]. The person detection algorithm based on deep learning method will reduce the time overhead of ranking surveillance video data and improve the efficiency.

In this paper, we aim to detect accurately and quickly the location information of people entering the area and their corresponding identity types for gas facilities task scenarios. We propose an improved Yolov4-tiny person detection model to intelligently monitor people entering the gas facilities to ensure proper operation of the gas facilities. The main contributions are as follows.

1) Based on the original two feature layers, a large-scale feature layer is added and the size of the anchor box is adjusted to improve the model's ability to detect small targets.

2) The attention mechanism module was introduced. By training to decide the feature fusion weights at different scales, the model can focus more on important features and be more accurate in feature extraction to improve the fine-grained detection effect of the model.

The paper is organized as follows. The Improved-Yolov4-tiny network is described in Section 2. The experiment process and the analysis of the results are described in section 3. Finally, Section 4 is the conclusion.

2. THE IMPROVED-YOLOV4-TINY MODEL
The idea of the Yolov4-tiny model is that the whole image is used as the input to the model network, and then the location of the target object bounding box and the corresponding class are regressed based on the anchor box. The main implementation method is to divide the input image into N x N grids, each grid is responsible for detecting the target whose center falls into that grid, and each grid corresponds to predict three pieces of information, the position of the target box and the confidence level.

In this section, we will focus on the model of the improved Yolov4-tiny. The detection of small targets has been improved by adding a large-scale feature layer. The attention mechanism is introduced to make the model more accurate for feature extraction and improve the fine-grained detection of the model. The network structure of the Improved-Yolov4-tiny algorithm is shown in Figure 1.
2.1 Multi-scale prediction
The original model network structure corresponds to a backbone feature extraction network containing 2 scale feature layers, 13×13×512 and 26×26×256 scale feature layers, respectively. In the paper, 52×52×128 large-scale feature layers are added to increase the effectiveness of the model for small target detection. So that the feature map with a smaller receptive field is responsible for the detection of small targets, and the feature map with a larger receptive field is responsible for the detection of large targets. The FPN structure is used to fuse feature layers of different scales to better understand the semantic features of images and to better accomplish the target detection task.

Small-scale feature layer: the feature layer size is 13×13×512, which is responsible for predicting larger targets, and the prediction output is 13×13×(n×7) through the convolution layer.

Mid-scale feature layer: the feature layer size is 26×26×256, which is responsible for predicting targets of moderate size, and the prediction output is 26×26×(n×7) through the convolutional layer.

Large-scale feature layer: the feature layer size is 52×52×128, which is responsible for predicting smaller targets, and the prediction output is 52×52×(n×7) through the convolution layer. Where n is the number of anchor boxes corresponding to this feature layer.

2.2 kmeans++ algorithm to generate anchor box
Since the object to be detected by the model is the person entering the gas facilities, and the size ratio of the person detection box is relatively fixed. In order to lighten the model, the number and size of the corresponding a priori boxes of each grid are adjusted. The original model contains 3 anchor boxes for each grid of each feature layer. For anchor box size generation, the size of all target boxes in the dataset is obtained and then clustered using the k-means algorithm for all size data.

The original model uses k-means algorithm to generate anchor box size, but k-means algorithm needs to determine the initial clustering center artificially, different initial clustering centers may lead to completely different clustering results, affecting the effect of anchor box size generation. The basic idea...
of k-means++ algorithm to select the initial clustering centers is that the initial clustering centers should be as far away from each other as possible, which is also in line with the logic of the clustering algorithm idea [15]. So, in this paper, we use k-means++ to generate the anchor box size. Define the distance by using the IoU function. IoU is known as the Intersection-over-union, which is calculated as the ratio of the intersection and the concatenation of two different target frames. The formula to define the clustering distance is shown in Equation (1).

\[ d = 1 - \text{IoU} \]  

(1)

The results obtained by clustering are used to set the corresponding size of anchor box on different size feature maps to speed up the training fitting of the model.

2.3. Introduction of attention mechanism multi-scale fusion

By extracting features through the feature extraction network, three feature layers of 13×13×512, 26×26×256 and 52×52×128 can be obtained. The pyramidal form of CNN hierarchical features is used to generate feature pyramids with strong semantic information at all scales simultaneously [16]. In order to utilize the features more effectively and process them efficiently, an attention mechanism module is introduced on the basis of the feature pyramid to enhance the detection of the model.

The attention mechanism module is introduced by adding spatial attention mechanism (SAM) and channel attention mechanism (CAM) between the feature layer and the feature pyramid to enhance the feature extraction ability of the model and increase the fine-grained detection effect of the model. The structure is shown in Figure 2.

The CAM module, whose main role is to add a weight to each channel of the feature layer to indicate the degree of association between that channel and the key information needed by the model. If the value of the weight is larger, the more important this corresponding channel is and the more the model needs to pay attention to it. Through the CAM, the model can pay more attention to strongly associated features and ignore or suppress weakly associated features, so that the model can accurately identify person's features, where the weights of each channel are trained by the model. The CAM is calculated as shown in Equations (2) and (3).

\[ W_c(F) = \sigma(FC(AVGPool(F))) + FC(AVGPool(F)) \]  

(2)

\[ F = F \times W_c(F) \]  

(3)

Where \( W_c(F) \) is the weight corresponding to the different channels of the feature layer. \( \sigma \) is the activation function sigmoid, \( FC \) is the fully connected layer, \( AVGPool \) is the global average pooling, \( MAXPool \) is the global maximum pooling, \( F \) is the feature layer.

The SAM module, whose main role is to add weights to the spatial locations of the feature maps. The feature maps of each channel correspond to the same spatial weights. The SAM, which pays more attention to the spatial information of the feature map, enables the model to learn to pay attention to the spatial location information of the feature map. Complementing the location relationship information that cannot be better obtained by the CAM, and then used to filter the individual location feature values in the feature map to highlight the applicable person features. The SAM is calculated as shown in Equations (4) and (5).
\[ W_s(F) = \sigma(Conv(AVGPool(F), MAXPool(F))) \]  

(4)

\[ F = F \times W_s(F) \]  

(5)

Where \( W_s(F) \) is the weight corresponding to the spatial location of the feature layer, directing the model to focus on the spatial features and the location relationship information obtained. The \( Conv \) are convolution layers.

The attention mechanism module is introduced by adding a SAM and a CAM between the feature layer and the feature pyramid. And add residual connections to each attention mechanism module. The core calculation formula is shown in Equation (6).

\[ F_n = F_n + F_n W_{c_n}(F_n) + (F_n + F_n W_{c_n}(F_n)) W_{s_n}(F_n + F_n W_{s_n}(F_n)) \]  

(6)

\( F_n \) is the \( n \)th feature layer, \( W_{c_n} \) is the \( n \)th channel attention weight, \( W_{s_n} \) is the \( n \)th spatial attention weight, and \( n \in [1,2,3] \).

3. EXPERIMENTAL PROCESS AND RESULTS ANALYSIS

3.1. Dataset

The data set consists of own dataset and public dataset, where the own dataset is acquired by the gas facilities monitoring cameras, and the public dataset using Pascal VOC2012. The data set can divide into 3 categories, Class A: only workers. Class B: only non-workers. Class C: workers and non-workers. The dataset contains a total of 9350 images and is randomly partitioned into a training set, a validation set and a test set. The overall situation of the dataset is shown in Table 1.

| Dataset    | Images | Worker | Non-worker | Person total |
|------------|--------|--------|------------|--------------|
| Train set  | 7350   | 4324   | 3982       | 8306         |
| Validation set | 1000   | 672    | 595        | 1267         |
| Test set   | 1000   | 624    | 513        | 1137         |
| Total      | 9350   | 5620   | 5090       | 10710        |

In order to increase the generalization ability of the model to reduce the effect of overfitting, image enhancement is performed on the dataset during the training phase of the model, mainly by random rotation, flipping, etc.

3.2. Experimental Progress

The model training in this paper was run on an Ubuntu 18.04 system with the Pytorch deep learning framework, using a GTX1080 GPU to accelerate the network training. The dataset and network model introduced above are used to detect and identify persons entering the gas facilities and to obtain detection results and verify the effectiveness of the method.

We use the pre-trained weights of Yolov4-tiny on the coco dataset as the initialization parameters of the network. The anchor box is generated by kmeans++ clustering method using a priori knowledge. The model training process was iterated for 200 epochs, and the Adam optimization algorithm was used to optimize the network. The learning rate is attenuated by the CosineAnnealingLR algorithm.

The target boxes of the dataset are clustered by Kmeans++ to obtain the anchor box. Figure 3 shows the accuracy comparison of kmeans algorithm and kmeans++ algorithm for generating anchor box under the selection of 1-10 clustering centers. Here the accuracy refers to the maximum IoU mean of each target box and anchor box.
Figure 3. Comparison of kmeans and kmeans++ clustering results

Observing Figure 3, we can find that after $k = 6$, the change of Kacc tends to be smooth, and for $k=6$, the Kmean++ algorithm obtained better results than the Kmeans algorithm, with an average accuracy (average IoU) of 0.829. Therefore, the clustering results under 6 clustering centers using Kmeans++ algorithm. The 6 anchor boxes, corresponding to 3 feature layers of different scales, each feature layer corresponds to two anchor boxes. The assignment of anchor box is shown in Table 2

| Feature layer | Anchor box            |
|---------------|-----------------------|
| 52 × 52       | (39, 179), (51, 119)  |
| 26 × 26       | (60, 194), (63, 238)  |
| 13 × 13       | (81, 141), (84, 244)  |

The model training is divided into two parts. Firstly, freezing the backbone features to extract the network parameters for training. The learning rate is set to 1e-3 and the learning rate is decayed by the CosineAnnealingLR algorithm with batch set to 64. Then unfreeze the backbone feature extraction network and perform the parameter adjustment of the whole network. The learning rate is set to 1e-4, and the learning rate is also decayed by the CosineAnnealingLR algorithm, with a batch of 64. The whole model training is completed in about 3h, and the changes of the loss values during the model training are shown in Figure 4.

Figure 4. Training loss variation

As we can see in Figure 3, the loss value gradually decreases as the network is trained, and the loss value stabilizes after the 175th epoch. The model finally reaches convergence.

3.3. Experimental result Analysis

To test the effect of the model, we conducted the following experiments to test the effect of different improvements on the model, and the experimental setup is shown in Table 3.
TABLE III. DIFFERENT IMPROVEMENT METHODS

| Method                      | Multi-scale | CAM | SAM |
|-----------------------------|-------------|-----|-----|
| Yolov4-tiny                 |             |     |     |
| Yolov4-tiny-M               | √           |     |     |
| Yolov4-tiny-M-CA            | √           | √   |     |
| Yolov4-tiny-M-SA            | √           |     | √   |
| Yolov4-tiny-M-CSA           | √           | √   | √   |
| (Improved-Yolov4-tiny)      |             |     |     |

Among them, Multi-scale is to add a large-scale feature layer and adjust the anchor box. CAM is the module that introduces the channel attention mechanism and SAM is the module that introduces the spatial attention mechanism. The experimental results are shown in Figure 5 and Figure 6.

![Figure 5. The mAP of different improvement methods](image)

![Figure 6. The FPS of different improvement methods](image)

FPS is the frame rate test result with Intel(R) Core (TM) i7-9750H CPU and GTX1650 GPU configuration.

As can be seen in Figures 5 and 6, the model with the addition of a large-scale feature layer (Yolov4-Tiny-M) improves the performance metric mAP by 1.4% compared to the original model (Yolov4-Tiny). After adding the attention mechanism module to the multi-scale feature fusion part, the performance metric mAP is all improved. When the multi-scale feature fusion module adds both spatial attention mechanism and channel attention mechanism, the model Improved-Yolov4-tiny improves the performance metric mAP by 3.1% compared with the original model. In the model performance metric FPS, the speed is slightly reduced compared to the original model due to the increased structure.
From the above, the Improved-Yolov4-tiny gas facilities person detection and identification model meets the requirements of the gas facilities person detection and identification monitoring task in terms of speed and accuracy. The detection effect is shown in Figure 7 and 8.

4. CONCLUSIONS
In this paper, we propose an Improved-Yolov4-tiny method that can be used to monitor people entering the gas facilities in real time and determine whether the type of person is worker or non-worker based on characteristics such as clothing. The model improves the detection of small targets by adding large-scale feature layers. The introduction of attention mechanisms for multi-scale feature fusion increases the effectiveness of fine-grained detection of people. The Improved-Yolov4-tiny detection algorithm not only meets the requirements of real-time monitoring, but also achieves better results at different resolutions. The model meets the requirements of gas facilities person detection and identification monitoring tasks in terms of speed and accuracy, and is of great significance for intelligent monitoring and securing the safety of gas transportation.

REFERENCES
[1] Chen X Y, Yan J W, Xiang Z. Research and Suggestions on Safety Management of Urban Gas Pipeline Network[J]. Chemical Enterprise Management, 2019.
[2] Viola P, Jones M J, Snow D. Detecting Pedestrians Using Patterns of Motion and Appearance[J]. International Journal of Computer Vision, 2005, 63(2):153-161.
[3] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. IEEE, 2005.
[4] Wang X, Han T X, Yan S. An HOG-LBP human detector with partial occlusion handling[C]// IEEE International Conference on Computer Vision. IEEE, 2009.
[5] Lowe D. Distinctive image features from scaleinvariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91–110, Nov. 2004.
[6] Xie Y, Zhang J, Li T, Liu, J. Video Surveillance Based Subway Construction Unsafe Behavior Detection and Warning. Journal of Huazhong University of Science and Technology Nature Science Edition, 2019, 47(10):46-51.

[7] Yang S. Mine personnel identification and detection based on faster RCNN. Information Recording Materials, 2020, 21(12):236-238.

[8] Ivasic-Kos M, Kristo M, Pobar M. Person Detection in Thermal Videos Using YOLO[J]. 2020.

[9] Bochkovskiy A, Wang C, Liao H. Yolov4: Optimal Speed and Accuracy of Object Detection[J]. 2020.

[10] Zhong Z, Zheng L, Kang G, Li s and Yang Y. Random Erasing Data Augmentation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 34(7).

[11] Shorten C, Khoshgoftaar T M. A survey on Image Data Augmentation for Deep Learning[J]. Journal of Big Data, 2019, 6(1).

[12] He K, Zhang X, Ren S and Sun J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 2015, 37(9):1904-1916.

[13] Sun X, Hao H, Liu Y, Zhao Y and Wang Y. Application of Yolov4 in Power Inspection Target Detection [J]. Modern Information Technology, 2020, 4(20):115-117.

[14] Oltean G, Florea C, Orghidan R and Oltean V. Towards Real Time Vehicle Counting using YOLO-Tiny and Fast Motion Estimation[C]// 2019 IEEE 25th International Symposium for Design and Technology in Electronic Packaging (SIITME). IEEE, 2019.

[15] Kubo Y, Nii M, Muto T and Kobashi S. Artificial humeral head modeling using Kmeans++ clustering and PCA[C]// 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech). IEEE, 2020.

[16] Lin T, Dollar P, Girshick R, He K, Hariharan B and Belongie S. Feature Pyramid Networks for Object Detection[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2017.