Object Recognition Through UAV Observations Based on Yolo and Generative Adversarial Network

Bo Li[1,2(✉)], Zhigang Gan[1,3], Evgeny Sergeevich Neretin[3], and Zhipeng Yang[1]

[1] School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
libo803@nwpu.edu.cn
[2] CETC Key Laboratory of Data Link Technology, Xi'an, China
[3] Moscow Aviation Institute, Moscow, Russian Federation

Abstract. Aiming at the object recognition through UAV, an intelligent object recognition model based on YOLO and Generative adversarial network is proposed in this paper. Firstly, the solution is given, and an object recognition model that can realize intelligent recognition is established. Then, in order to improve the resolution of the identified images, an image resolution enhancement model based on generative adversarial networks is built. After that, the structure and parameters of the recognition model and image resolution enhancement model are adjusted through the simulation experiments to improve the accuracy and robustness of the object recognition. Finally, the object recognition model based on YOLO and generative adversarial network in this paper is verified through UAV.

Keywords: UAV · Machine learning · Object recognition

1 Introduction

In recent years, UAVs have widely used in different types of civilian missions, such as target tracking, wildlife protection, disaster rescue and 3D reconstruction. In view of the rapid development of advanced UAV technology, target tracking is an important research and application direction of UAV. Object recognition technology has huge requirements in areas such as autonomous UAV-guided landing, criminal vehicle or personnel hunting, military target reconnaissance and strike. However, the above tasks have high requirements on the accuracy, stability, and robustness of object recognition. Traditional object recognition can no longer meet the current task requirements. In order to complete the precise recognition tasks, today's more advanced technology must be used. At the same time, there are many difficulties in applying object recognition. For example, complex lighting changes, occlusion between targets and scenes, similar interference in the background, and camera perturbations all make target tracking tasks more difficult.

B. Li et al. (Eds.): IoTaaS 2020, LNICST 346, pp. 439–449, 2021.
https://doi.org/10.1007/978-3-030-67514-1_35

The advantage of the traditional object recognition algorithm is that it can clearly display the recognition features, and if the object is simple, the outline is clear, or the color contrast is sharp, then the traditional object recognition algorithm can be used to identify the target well. However, deep learning methods work well for high-dimensional data. Without manual selection of features, classification results with good effect can be obtained by deep learning methods. Among the object recognition algorithms based on deep learning, models such as the YOLO(You only look once) [1, 4] model and R-CNN have achieved good recognition results. YOLO can detect video very fast basic YOLO model processes images in real-time at 45 frames per second.

Considering the YOLO model's efficient and accurate recognition ability for video images, this article decided to use YOLO to complete object recognition missions. Considering the limited sharpness of object images, this paper uses image enhancement technology based on GAN (Generative adversarial network) [2] to assist object recognition. In terms of image resolution enhancement technology, the traditional problem is that when the magnification of the image resolution is more than 4 times, high-frequency information is missing, and the realism in details is missing. However, SRGAN (Super-resolution generative adversarial network) [3] model, an image resolution enhancement method based on generation adversarial network, can better generate image details when dealing with the enhancement problem with larger resolution magnification.

In this paper, we have combined YOLO and GAN in object recognition to ensure the stability and efficiency of video images recognition and the clarity of UAV recognition images.

The contributions of this paper are summarized as follows:

1. We combined YOLO and SRGAN to propose a new method for object recognition through UAV observations. UAV realizes object recognition through YOLO and enhances the resolution of the recognition images through SRGAN, which can realize the high-quality and fast object recognition.
2. We use YOLO algorithm to achieve object recognition and SRGAN algorithm to enhance the resolution of the recognition image, which can achieve high-quality rapid detection of the target by the UAV.
3. We constructed a UAV platform and verified the effectiveness of our proposed methods on the UAV platform. Through a series of experiments, we show that the UAV can quickly and efficiently complete object recognition and realize the resolution enhancement of the recognized images.

The reminder of this paper is organized as follows. In Sect. 2, this article defines a solution for completing object recognition on UAV, and details the implementation of object recognition and image resolution enhancement. In Sect. 3, this paper trains the recognition model and completes the test on the UAV. Finally, conclusions are presented in Sect. 4.

2 Problem Formulation and Learning Algorithm

2.1 Problem Description

The research body is divided into two parts: UAV and ground laptop. The height advantage and mobile advantage of UAV are utilized to complete the acquisition of ground images, and the acquired images are transmitted to the ground laptop in real time through the image transmission system of UAV [6–9]. The ground laptop completes the object recognition of interest and performs resolution enhancement operations on part of the bounding box of image [10–12]. The process of UAV object recognition is shown in Fig. 1.

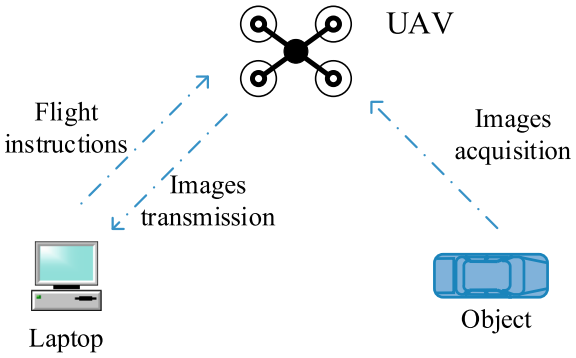

Fig. 1. The process of UAV object recognition

The object recognition process of this paper is as follows. UAV collects images from the ground through the camera and transmits the images to the ground laptop in a real time. The real-time images are input into the YOLO network to complete the object recognition [13–15]. The recognition images are enhanced by generative adversarial network to improve the recognition of object by human eyes. The implementation process is shown in Fig. 2.

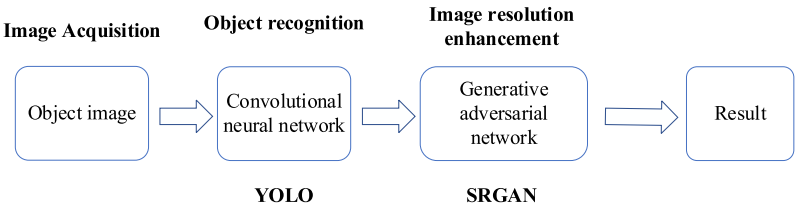

Fig. 2. The implementation process of recognition

2.2 Object Recognition Based on YOLO

Most of the current object recognition algorithms treat the object recognition problems as classification problems. However, the YOLO model abstracts object recognition as regression problems, directly from image pixel input to output boundary box and category probabilities. YOLO is an object recognition system based on a single neural network proposed by Joseph Redmon and Ali Farhadi et al. in the paper You Only Look Once: Unified, real-time Object Detection in 2015 [1].

YOLO is based on the convolutional neural networks to achieve the recognition. YOLO consists of 24 convolutional layers [5] and 2 fully connected layers. The neural networks extract features from the images through the convolution layers, and uses fully connection layers to predict the output probability and the position information of the bounding box.

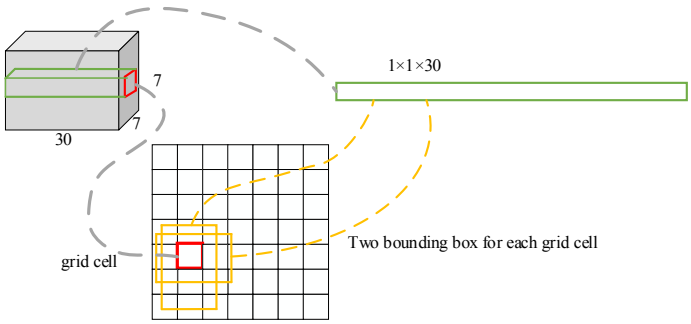

Fig. 3. The process of object recognition by YOLO (Color figure online)

In the YOLO network, the neural network predicts the location and confidence of each bounding box. Confidences includes the confidence of the object in the bounding box and the accuracy of the position prediction of the bounding box. When an image is input into the YOLO, the output is a $7 \times 7 \times 30$ three-dimensional tensor. Figure 2 shows the process of object recognition by YOLO.

As shown in Fig. 3, the red box represents the grid cell, and the yellow boxes represent the bounding boxes. The input image is divided into 7×7 grid cells, each grid cell corresponds to two output bounding boxes predicted by YOLO. 30 in the output tensor represents the position and category of each bounding box.

The loss function of YOLO is defined as (1), which is divided into two parts: position error and classification error. Classification errors include confidence error and probability error. Confidence errors are divided into confidence error with object and confidence error without object.

$$
\begin{aligned}
loss = {} & \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] + \\
& \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} [(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2] \\
& + \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (C_i - \hat{C}_i)^2 + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (C_i - \hat{C}_i)^2 \\
& + \sum_{i=0}^{S^2} 1_i^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2
\end{aligned}
\tag{1}
$$

where, λ_{coord}, λ_{noobj} respectively represent the weights of position information and category information defined in the loss function, i represents the i-th ($i = 0, \ldots, S^2$) grid cell, j represents the j-th ($j = 0, \ldots, B$) bounding box, 1_{ij}^{obj} represents j-th bounding box in grid cell i to be predict the object, 1_i^{obj} indicates that the object appears in grid cell i, $(\hat{x}_i, \hat{y}_i)$ represents the center position information of the prediction object position, $(\hat{w}_i, \hat{h}_i)$ represent the width and height of the prediction recognition box, $\hat{C}_i$ represents the classification category for object i, $\hat{p}_i(\mathrm{c})$ represents the confidence for category.

2.3 Image Resolution Enhancement Based on SRGAN

Since the implementation and test platform of this paper is the UAV, the images collected by the UAV are input into the YOLO model to complete the object recognition mission. Considering that the UAV is at high altitude, the clarity of the image information collected from object is limited. Therefore, after the object recognition is completed, the resolution enhancement of recognition image is carried out to improve the recognition effect. This paper uses Generative Adversarial Network (GAN) [2] which has excellent performance in generating images and enhancing image resolution.

In order to enhance the image sharpness of images, this paper decided to use SRGAN (Super-Resolution Using a Generative Adversarial Network), a branch model of generation antagonism network, to achieve image resolution enhancement. SRGAN recovers high-frequency information of images by Perceptual loss and Adversarial loss. SRGAN makes the generated image and target image more similar in style by comparing the features extracted from the convolutional neural network and the features extracted from the target image through the neural network. SRGAN consists of a generation network and a discrimination network. The generation network consists of 5 convolutional layers and 5 layers of residual networks. Convolutional layers are used to extract image features. The residual network can promote the training effect and solve the problems of gradient disappearance and gradient explosion. The discrimination network consists of 4 convolutional layers for extracting input image features. The generation work takes low-resolution images as input and tries to generate high-resolution image data. The discrimination network takes the real high-resolution image

and the image generated by the generation network as input, and predicts whether the current input comes from real data or generated image.

The loss function of generation network is defined as, which consists of (3), (4), (5). $g_{contentloss}$ represents the content loss of the generated image, $g_{VGGloss}$ represents the loss after feature extraction, $g_{adversarial}$ represents the training loss of the generation network. (x, y) represents the image pixel coordinates, and r_w, r_H represent the width and height of the image. $\phi_{i,j}(I^{HR})_{x,y}$ is the output of the high-resolution image after the feature extraction, $\phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y}$ is the output of the generated image after the feature extraction.

$$g_{loss} = g_{contentloss} + g_{VGGloss} + g_{adversarial} \tag{2}$$

$$g_{contentloss} = \frac{1}{rWrH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \tag{3}$$

$$g_{VGGloss} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2 \tag{4}$$

$$g_{adversarial} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR})_{x,y}) \tag{5}$$

The loss function of discrimination network is defined as

$$E_{I^{HR} \sim p_{train}}(I^{HR})[log\, D_{\theta_D}(I^{HR})] + E_{I^{LR} \sim p_G}(I^{LR})[\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})_{x,y}))] \tag{6}$$

where I^{HR}_{xy}, I^{HR} represent high-resolution images, I^{LR} represent low-resolution images, and $G_{\theta_G}(I^{LR})_{x,y}$ represents the generation results of the generation network. $D_{\theta_D}(I^{HR})$ represents the discriminant result of the discrimination network with high-resolution images as input, $D_{\theta_D}(G_{\theta_G}(I^{LR})_{x,y}$ is the result of the discrimination network with generated images as input, $I^{HR} \sim p_{train}$ is the high-resolution images from the training data set, and $I^{LR} \sim p_G$ represents the generated images from generation network.

When training the generation adversarial network, alternately train the generation network and the discrimination network. Within a period of time, the parameters of the generation network are fixed and the discrimination network is optimized. In the next period of time, the parameters in the discrimination network are fixed and the generation network is optimized. The training process of SRGAN is shown in Fig. 4.

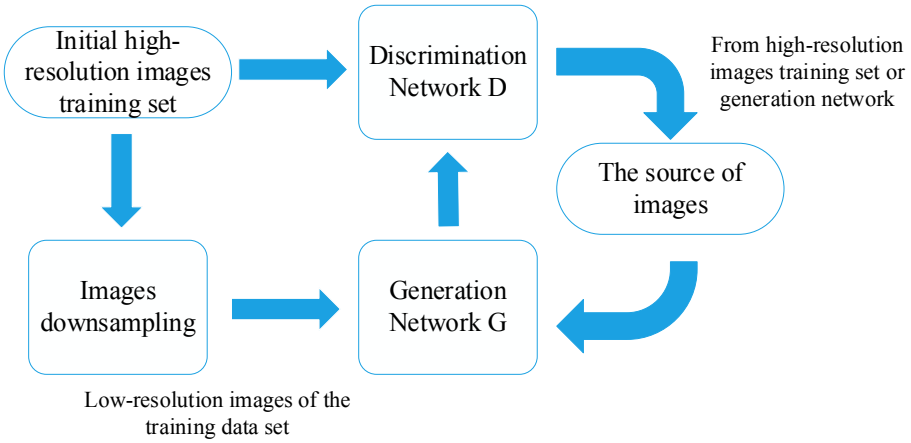

Fig. 4. The training process of SRGAN

The image resolution enhancement algorithm is summarized in Table 1.

Table 1. SRGAN algorithm

Algorithm SRGAN
In generation network G and discrimination network D
for number of training iterations **do**
for k steps **do**
Sample minibatch images from high-resolution images $\{x^{(1)}, x^{(2)}, x^{(3)}, ..., x^{(m)}\}$
Sample minibatch images from low-resolution images $\{z^{(1)}, z^{(2)}, z^{(3)}, ..., z^{(m)}\}$
Update the discrimination network by ascending its stochastic gradient: $$V = \frac{1}{m}\sum_{i=1}^{m} \log D(x^i) + \frac{1}{m}\sum_{i=1}^{m} \log(1 - D(G(z^i)))$$
end for
Sample minibatch images from low-resolution images $\{z^{(1)}, z^{(2)}, z^{(3)}, ..., z^{(m)}\}$
Update the generation network by ascending its stochastic gradient.
end for

3 Experiment and Analysis

3.1 The Object Recognition Experiment by YOLO

This paper completes the construction and trains the recognition model based on TensorFlow module.

The changes in the loss of the recognition model training process are shown in Fig. 5. It can be seen that the recognition result reaches the expected value after 1,000,000 training steps. The recognition rate of the final recognition model on the test set reached about 86%.

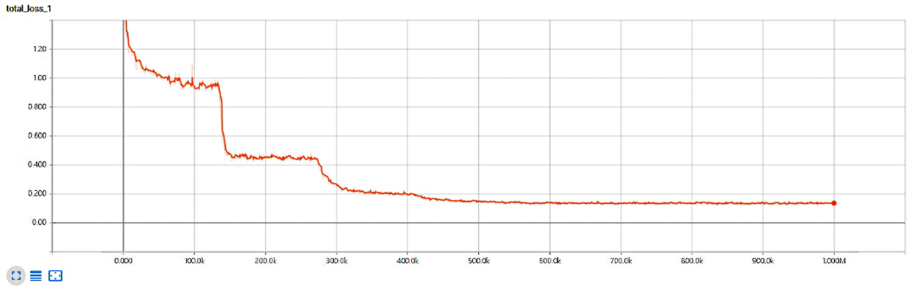

Fig. 5. The loss function during training

When testing on UAV: The experimental procedure is that in the case of low-speed movement of the unmanned vehicle, the UAV is manually operated to take off and fly in the direction of the object. During this period, the camera on the UAV has been collecting the ground images and transmitting them to the ground laptop in real time. While receiving the images, the trained YOLO model has been running. When the unmanned vehicle appears on the screen of the laptop, the bounding box is superimposed on the real-time video, which means the object is found. In order to detect the robustness of the recognition algorithms, pedestrians are introduced as interference during the test.

According to the Fig. 6, it can be seen that the YOLO model can accurately complete the recognition task of the unmanned vehicle moving at low speed at various heights, positions and angles.

3.2 The Resolution Enhancement Experiment by SRGAN

In this paper, the high-resolution image data sets are composed of images similar to the target texture features, while the low-resolution image data sets are obtained from the down-sampling of the high-resolution image data sets.

The loss of generation networks and discrimination networks are shown as Fig. 7. The left graph is the loss of generation network, and the right graph is the loss of discrimination network.

Fig. 6. The recognition results on UAV

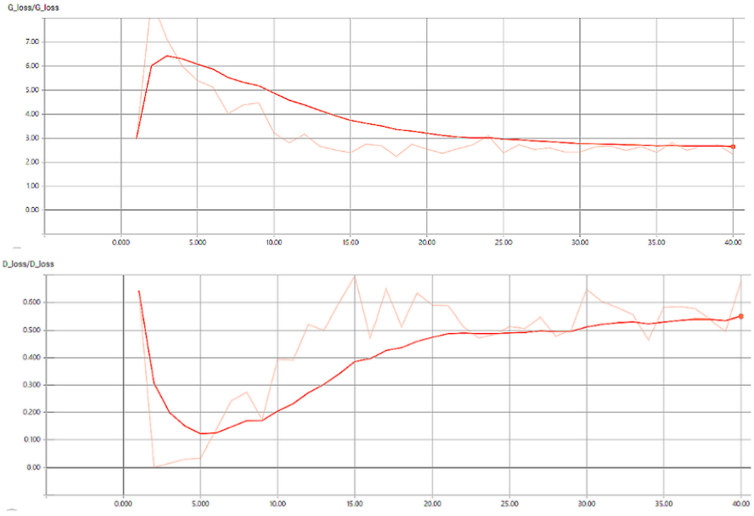

Fig. 7. The loss function during SRGAN training

Due to the poor definition of the image information collected by UAV, SRGAN was used to enhance the resolution after completing the recognition tasks. The results are shown in Fig. 8.

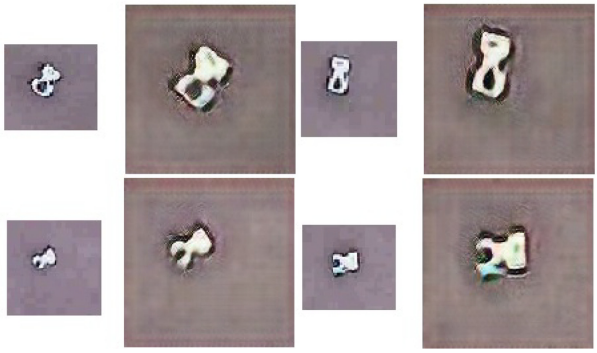

Fig. 8. The results of image resolution enhancement

From the experimental results in Fig. 8, it can be seen intuitively that the sharpness of the unmanned vehicle has been enhanced, which shows that the use of SRGAN technology in UAV platform target detection is beneficial to human observation.

4 Conclusion

This paper proposes a recognition model based on YOLO and generative adversarial network for UAV object recognition. First, we construct a real-time recognition model based on YOLO. Then, in order to improve the resolution of the images, we construct an image resolution enhancement model based on generative adversarial network. Through a series of experiments, we adjust the structure and parameters of the recognition model and image resolution enhancement model to improve the accuracy and robustness of object recognition. Finally, we carry out experiments on the UAV platform, which verify the effectiveness of this paper based on YOLO and generative adversarial network.

Acknowledgments. This research was supported by The Aeronautical Science Foundation of China (No. 2017ZC53021) and The Open Project Fund of CETC Key Laboratory of Data Link Technology No. CLDL-20182101).

References

1. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)
2. Goodfellow, I.J., et al.: Generative Adversarial Networks, June 2014
3. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4681–4690 (2017)

4. Simon, M., Milz, S., Amende, K., Gross, H.-M.: Complex-YOLO: real-time 3D Object detection on point clouds, September 2018
5. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: International Conference of Learning Representation (ICLR), January 2016
6. Zeng, F., Shi, H., Wang, H.: The object recognition and adaptive threshold selection in the vision system for landing an Unmanned Aerial Vehicle. In: 2009 International Conference on Information and Automation, pp. 117–122 (2009)
7. Kechagias-Stamatis, O., Aouf, N., Nam, D.: 3D automatic target recognition for UAV platforms. In: 2017 Sensor Signal Processing for Defence Conference (SSPD), London, pp. 1–5 (2017)
8. Ibrahim, A.W.N., Ching, P.W., Seet, G.L.G., Lau, W.S.M., Czajewski, W.: Moving objects detection and tracking framework for UAV-based surveillance. In: 2010 Fourth Pacific-Rim Symposium on Image and Video Technology, Singapore, Singapore, pp. 456–461 (2010)
9. Wang, C., Zhao, R., Yang, X., Wu, Q.: Research of UAV target detection and flight control based on deep learning. In: 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, pp. 170–174 (2018)
10. Li, X., Yan, B., Wang, H., Luo, X., Yang, Q., Yan, W.: Corner detection based target tracking and recognition for UAV-based patrolling system. In: 2016 IEEE International Conference on Information and Automation (ICIA), Ningbo, China, pp. 282–286 (2016)
11. Sommer, L., Schumann, A., Muller, T., Schuchert, T., Beyerer, J.: Flying object detection for automatic UAV recognition. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, pp. 1–6 (2017)
12. Wang, J., Pundit, S.P., Abdelzaher, A.F., Watts, M.: Asynchronous localization of ground objects using a 2-UAV system. In: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, pp. 316–321 (2019)
13. Zientara, P.A., Choi, J., Sampson, J., Narayanan, V.: Drones as collaborative sensors for image recognition. In: 2018 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, pp. 1–4 (2018)
14. Zhe, L.I., Jianzeng, L.I.: Analysis of blurred image restoration for small UAV. J. Ordnance Equipment Eng. **40**(3), 165–168 (2019)
15. Yang, C., Chen, D., Liu, Z., et al.: Image dehazing method based on deep learning. J. Ordnance Equipment Eng. **40**(10), 131–135 (2019)