Research Article

Benchmarking the Robustness of Object Detection Based on Near-Real Military Scenes

Yue Zhang , Long Ye , Li Fang , Wei Zhong , Fei Hu , and Qin Zhang

Key Laboratory of Media Audio & Video (Communication University of China), Ministry of Education, Beijing 100024, China

Correspondence should be addressed to Long Ye; yelong@cuc.edu.cn

Received 4 October 2021; Revised 9 January 2022; Accepted 8 February 2022; Published 5 March 2022

Academic Editor: Yushu Zhang

Copyright © 2022 Yue Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

According to the technical requirements of intelligent development of auxiliary combat system, we construct a visual intelligent test platform. A near-real military scene dataset based on physical rendering is built, which contains 11,000 remote sensing images collected by an analog camera taking pictures in different illumination, weather environment, camera shooting angle, and scene scale condition. Besides, we add a natural style transfer module for a single unmodeled military scene image's multienvironment generation. We conduct experiments to evaluate the stability of several UAV remote sensing image object detection algorithms. Based on the quality and speed value of the tested algorithms, the adaptability scores in different environments are calculated. Furthermore, we propose a comprehensive evaluation index system of military remote object detection based on a hierarchical model. We envision that our comprehensive benchmark will play a role in the evaluation of algorithm capability for military object detection tasks and the improvement of training algorithm capability.

1. Introduction

In terms of military applications, compared with intelligent object recognition tests in real scenes, artificial near-real virtual scenes have more significant advantages on cost, fidelity, repetition rate, and controllability. First, in order to meet the technical requirements of the intelligent development of auxiliary combat systems, we constructed a visual intelligent test platform. Compared with the real scene, the virtual platform can arbitrarily transform the position and scale of the objects in the virtual scene without changing the relevant restrictions such as the details of the target. Second, inspired by many style transfer approaches based on GAN [1–4], we proposed an algorithm, which included a generator to render natural military scene images in different weather in the same camera view based on CycleGAN [5], so as to provide specific datasets (In real military missions, pictures of different weather can be generated in real time for selected scene pictures by the generator trained by images collected on our intelligent platform. We only make a brief of the algorithm and do not show the dataset content

specifically.) for other military missions, and improved the generalization ability of intelligent algorithms.

Images can be generated from the natural phenomenon generation module of physical rendering by Unreal Engine 4. For example, cloud rendering is simulated by using a ray projection algorithm based on image sequence in a three-dimensional scene [6–11]. We get the large-scale images of the virtual scene in different environment based on the physics engine. From the perspective of expression style and content characteristics of the deep convolutional neural network, the natural style generation of a single military scene image was realized. For an unmodeled single real military scene image, the generator of our style transfer algorithm can render it.

Based on the intelligent platform, we can measure and evaluate an intelligent detection algorithm in multiple conditions such as different illumination, environment, scale, and angle. The algorithm to be evaluated was applied to different scene images, in which the light intensity was divided into bright light, moderate intensity light, and weak light, and the weather conditions include sunny, rainy, snowy,

and foggy. We also used different camera capturing view including distant-view, mid-view, and close-view to simulate the scales of images. Then, the scene image's angle was divided to three. The mAP and FPS scores evaluated by nonadaptive system can only be evaluated on a single fixed dataset. On the platform of near-real virtual scene, we can measure image detection methods from illumination, environment, perspective, scale adaptability, and other aspects to obtain the performance score and speed score under different conditions. We completed semantic annotation of basic scenes and used these annotation and context conversion technology to generate a large number of near-real training pictures. Compared with the existing 2D images [12] and 3D indoor scenes [13–15], the 3D environment of this project has the characteristics of more complex environment and stronger real simulation.

For intelligent algorithms on different missions, after scoring on all four indicators, we added a comprehensive evaluation model based on the analytic hierarchy process [16]. Firstly, through questionnaire surveys and investigation of a large number of application scenarios of military target monitoring, the importance comparisons among the four evaluation indicators were obtained, and the qualitative judgment difficult to be quantified was transformed into an operable importance comparison. The evaluation indexes were compared pairwise to get the judgment matrix, and the consistency test was carried out. Then, the weight of the judgment matrix was obtained by the arithmetic average method, geometric average method, and eigenvalue method. We put the obtained light adaptability score S_l , environmental adaptability score S_e , scale adaptability score S_s , and angle adaptability score S_a to calculate the index value into the formula of the illuminant index, the environment index, the scale index, and the angle index, respectively. Finally, combining the weight of the judgment matrix, the comprehensive evaluation score of the algorithm was obtained.

2. Related Work

Visual intelligent recognition datasets, ImageNet [17], MSCOCO [18], PASCAL VOC [19], KITTI [20], and Oxford Robot car [21], promoted the development of visual intelligent technologies such as face detection and recognition, pedestrian tracking, gait recognition, visual question answering, and scene understanding. HiEve [22] played a good role in multiobject tracking, attitude estimation and tracking, and motion recognition. CrowdHuman [23] focused on challenging scenarios under a variety of crowded and complex events whereas previous datasets were mostly associated with normal or relatively simple scenarios. Cityscapes [24] referred to the semantic understanding of urban street view, aiming to apply the evaluation visual algorithm to semantic segmentation tasks to better support the object detection tasks on urban scene. However, due to the limitations of filming equipment and scenes, there is still a gap between the dataset collected from the real environment and the requirements of simulating military application-related scenes. Gaidon et al. [25] recorded a video dataset in a virtual environment to extend KITTI for analyzing the

multiobject tracking capability of algorithms. Qiao et al. [26] constructed a virtual supermarket environment and collected images containing objects of different scale to analyze the robustness of object detection algorithms to scale changes. Gupta et al. [27] evaluated their proposed visual navigation algorithm by measuring task completion rate and average search path in a highly realistic virtual environment. Xia et al. [28] proposed a large-scale dataset for Object Detection in Aerial Images, DOTA, whose rotation label and polygon label were very helpful for the identification of some categories, such as ships. But the dataset still has a problem of limited remote sensing data. At the same time, some AI research teams have opened a large number of virtual environments for AI training and testing, such as AirSim, OpenAI, and DeepMind, striving for the commanding heights of AI research and testing platforms. On the other hand, although the datasets constructed by using virtual scenes has excellent performance, the cost is high, so how to transfer the style from a single image to obtain images in different environments is also worth exploring. Zhu et al. [5] proposed the CycleGAN model to realize the conversion of sample data without pairing, which can be applied to image style transfer tasks.

For remote sensing object detection tasks, it is extremely challenging due to the strong restriction of scene illumination, camera capturing angle, and scale of the captured images. Xu et al. [29] restored the low-illumination image through decomposition and enhancement, which greatly improved the target detection task under illumination changes. For objects with high density and small size and complex background, Long et al. [30] used a combination of a variety of traditional methods and depth methods to detect objects with small size, partial occlusion, and those not in the field of vision, as well as objects appearing in the dark background, which had a very good effect. Wang et al. [31] proposed MS-VANS to solve the multiscale problem in the target detection process by combining visual attention and loss function for data enhancement, aiming at target rotation and scaling and background clutter. Hendrycks and Dietterich [32] introduced a benchmark to evaluate the robustness of the recognition model for common corrosion images. Michaelis et al. [33] also proposed a benchmark to discuss the application of existing target detection algorithms in autonomous driving tasks. For comprehensive test indicators, Saaty [16] proposed an analytic hierarchy process (AHP), which combined qualitative and quantitative decisions with multiple plans or goals.

3. Near-Real Military Visual Intelligent Platform

In the whole framework of a visual task, the visual intelligent test platform plays a decisive role, which is of great significance to the improvement of the algorithm evaluation system and the construction of the visual task system. How to evaluate the visual detection algorithm is an important problem. In order to better evaluate the performance of the existing algorithms scientifically, we build a remote sensing image detection algorithm test platform on the established

near-real virtual scene dataset. The performance of the object detection model can be measured from four aspects: illumination, environment, angle, and scale adaptability. Combined with analytic hierarchy process (AHP), a comprehensive index evaluation model is proposed.

3.1. Establishment of the Dataset. To comprehensively test the recognition and tracking performance of the intelligent algorithm in different natural scenes, it is necessary to carry out real-time scene style conversion in the constructed near-real scenes to improve the generalization ability of the recognition algorithm. The natural phenomenon generation technology based on physical rendering needs to obtain the complete spatial information of 3D scene species, so it is difficult to process a single military scene image. To solve this problem, we adopt a multiway natural phenomenon scene generation route [34–37]. For 3D near-real scenes that have been fully modeled, physical particle rendering based on the UE4 software is used to complete scene conversion of natural phenomena such as rainfall, snowfall, and fog. For a single military scene image, images of various natural phenomena generated by the physics engine are used as the training set, and the style transfer model of natural scene phenomena is trained by deep learning technology. The flow chart is as follows (see Figure 1).

3.1.1. Physical Rendering Based on Natural Scene Generation Module. Unreal Engine 4 provides a powerful particle system, which can create a variety of complex visual effects. The physics rendering engine based on UE4 includes two rendering methods: particle rendering and volume rendering, which are, respectively, applied to the scene of rain and snow and cloud rendering.

- (i) *Rain and Snow Rendering Based on Particle System.* Making an n -dimensional vector $P^n = \{a_1, a_2, \dots, a_n \mid n \geq 3, n \in I\}$ represents a single particle, then define the mapping for a single particle to a positive integer set. Each particle gets a map; then, we make the mapping from I_t to P^n be $Q(t) = \{P^t : I_t \rightarrow P^n \mid I_t \subset J, n \geq 3, n \in I, t \in R\}$, and let $W(i) = P^n$ represents the state of the particle. The finite set of particle mapping is defined as a particle system, expressed as $S(t) = \{Q(t) \mid t \in \{t_0, t_1, \dots, t_m\}\}$, in which S is the set of particle system at different moment t_m . The state of the particle system at the initial moment is expressed as $S(t_0)$.

Based on the above particle system description method, by adjusting the shape and appearance, motion speed, and life cycle of the particle system, the style generation of natural phenomena such as rainfall and snowfall can be realized in near-real scenes.

- (ii) *Cloud Rendering Based on Volume Rendering.* It is necessary to use volume rendering algorithm to create a real and random cloud weather environment in virtual scene. At present, commonly used volume rendering algorithms include ray projection algo-

rithm, split-deformation algorithm, and frequency domain volume rendering algorithm. The ray projection algorithm uses a solution based on ray scanning, which is in line with common sense of human life, and the algorithm can be easily transplanted to GPU for implementation. Therefore, we use the ray projection algorithm to simulate the volume cloud and gas haze

Ray-casting algorithm is a direct volume rendering algorithm based on image sequence. From each pixel of the image, along a fixed direction, usually the gaze direction, it launches a light across the image sequence. In this process, the image sequence is sampled to obtain the color information, and the color values are accumulated according to the light absorption model until the light passes through the whole image sequence. The final color value is the color of the rendered image. The renderings based on the 3D engine are shown in Figure 2.

3.1.2. Natural Scene Generation Module Based on CycleGAN. The generation method based on physics engine can not only transform natural environment quickly and accurately for the nearly real scene with complete modeling but also use the generated images as training samples for the natural environment generation system of single military scene image. Based on the large-scale training dataset generated by the physics engine, this project, from the perspective of the expression style and content characteristics of the deep convolutional neural network, improves the generator based on CycleGAN algorithm and realizes an updated generator.

- (i) *Our Generator.* In this part, an improved generator composed of an encoder, a decoder, and a feature transfer module is designed. The structure of the feature transfer module is changed, and the image features are interpolated. The semantic attributes of the image are modified, and the learning of image features is realized successfully. Our baseline selects CycleGAN model, and the generator is composed of residual structural blocks and sampling convolution layers. Because the residual network is easy to train and has a strong capability of feature extraction, the application of this structure block has a good effect on image generation tasks. The network consists of two generators and two discriminators, which refers to the rainy day image as the domain X and the sunny day image as the domain Y , respectively. A mapping of a different domain is learned through a circular consistency process so that a model can be trained with only two types of images
- (ii) *Feature Transfer Module.* The generator of the original CycleGAN is composed of two downsampling convolution layers, nine residual modules, and two upsampling convolution layers. In the improved model, the generator of the two domains is integrated, and the depth feature transfer structure is added on the basis that the lower sampling layer, the upper sampling layer, and the residual module

FIGURE 1: Multipath generation system of natural scene images.

FIGURE 2: (a) Fog rendering based on volume rendering. (b) Cloud rendering based on volume rendering. (c) Alternate rendering effects day and night based on sunlight intensity.

remain unchanged. Under the condition that the 9 residual modules and the network depth remain unchanged, the two subsampling convolution layers with a step size of 2 and the first three modules of the residual module are used as encoders together, and the last three modules are used as decoders together with the two upsampling convolution layers with a step size of 1/2. The middle three residual modules are used as feature transfer modules from X domain to Y domain and from Y domain to X domain, respectively. The improved model adopted military scene dataset included images of sunny day and rainy day for training. The model adopted Adam optimizer, and the batch size was set as 1, and the initial learning rate was set as 0.0002. 200 epochs were trained, and the number of iterations was the number of images of a single category in the training set

3.2. Evaluation Index System of UAV Remote Sensing Image Detection Algorithm

3.2.1. Indicator Overview. Aerial images have the characteristics of scale diversity, perspective particularity, small target problem, multidirection problem, and high background complexity. The algorithm detection platform is built based on the near-real virtual scene dataset, which contains scoring module of illumination adaptability, environmental adaptability, scale adaptability, and angle adaptability. Through this platform, the evaluation scores of the algorithm model in illumination index, environment index, scale index, and angle index can be obtained.

- (i) *Illumination Index.* It is unreasonable to apply the radiometric correction coefficient obtained by the only reference object to all images due to the low flying altitude of micro-UAV and the constant changes of illumination and shooting angle during its flight. This will make continuous images have uneven illumination, resulting in inconsistent image saturation, brightness, and other problems. Illumination index mainly measures the performance of the algorithm model under different illumination intensity
- (ii) *Environmental Index.* Due to meteorological conditions or air pollution and other factors in some areas, there is a lot of haze, which seriously reduces the quality of UAV image acquisition and affects the subsequent image analysis and understanding. The environmental index mainly evaluates the performance of the algorithm model in different environmental scene, that is, the level of test accuracy
- (iii) *Scale Index.* The shooting height of aerial remote sensing images is between hundreds of meters and tens of thousands of meters. The size of similar detection targets on the ground is not the same, such as fighter jets, ships, and other military equipment, the size based on tens of meters to hundreds of meters. Scale index mainly evaluates

the performance of the algorithm model in different scale scenes

- (iv) *Angle Index.* Most remote sensing images are taken from a top view from high altitude, but most of the datasets commonly used by deep learning models are images from a ground-level perspective. In practical application scenarios, the angle shooting mode for the same target is variable, and the image detection learning model obtained only in the training of conventional datasets will have a worse effect when applied to the UAV remote sensing detection task in the military scenes. Angle index mainly evaluates the performance of the algorithm model in different angles

Each index is composed of a four-dimensional vector, and the calculation steps are shown as below. We enter the test platform with the evaluation algorithm model and apply the evaluation model to different lighting, environment, angle, and scale, respectively, lighting 1/environment 1/angle 1/scale 1, lighting 2/environment 2/angle 2/scale 2, ..., light n /environment n /angle n /scale n . Taking the illumination index as an example, the detection quality score and speed score of a single frame image are obtained under the scene of illumination n , and then, the mean and standard deviation of the quality and speed scores under the illumination is calculated. The mean value represents the performance of the algorithm, and the standard deviation represents the stability of the algorithm index. For example, compared with the traditional KITTI dataset, our dataset is more flexible to acquire (see Figure 3).

3.2.2. Quality Score and Speed Score. Assume that the quality score of the model to be evaluated under light/environment/angle/scale i ($1 \leq i \leq n$) is A_{ij} , which is the mAP value. The speed score is B_{ij} , which is the FPS value. The mean qe and standard deviation qe_v of quality score and the mean se and standard deviation se_v of speed score of the model to be evaluated are counted under all conditions:

$$\begin{aligned}
 qe &= \frac{\sum_{i=1}^n A_{ij}}{n}, \\
 se &= \frac{\sum_{i=1}^n B_{ij}}{n}, \\
 qe_v &= \sqrt{\frac{\sum_{i=1}^n (A_{ij} - qe)^2}{n}}, \\
 se_v &= \sqrt{\frac{\sum_{i=1}^n (B_{ij} - se)^2}{n}}.
 \end{aligned} \tag{1}$$

In A_{ij} , j can be 1, 2, 3, and 4. A_{i1} , A_{i2} , A_{i3} , and A_{i4} represent the mAP score of an algorithm under illumination i ($1 \leq i \leq n$), environment i , scale i , and angle i , respectively;

In B_{ij} , j can be 1, 2, 3, and 4. B_{i1} , B_{i2} , B_{i3} , and B_{i4} represent the FPS score of an algorithm under illumination i ($1 \leq i \leq n$), environment i , scale i , and angle i , respectively.

FIGURE 3: (a) KITTI: the view is limited to relatively fixed angle, and the conversion of vision is severely limited. (b) Our dataset: in our intelligent platform, images can be captured from different angle and scale.

3.2.3. Adaptability Score of Each Index. The calculation process of illumination index, environmental index, scale index, and angle index is shown in Figure 4. The algorithm to be evaluated is tested to obtain illumination index (qi, qi_v, si, si_v) of the model to be evaluated. The scoring formula of environmental index, scale index, and angle index is the same as the above formula, and the environmental index (qe, qe_v, se, se_v), scale index (qs, qs_v, ss, ss_v), and angle index (qa, qa_v, sa, sa_v) are obtained, respectively.

The expression method of the algorithm index is shown in Equation (2).

$$I_k = \left(\frac{\sum_{i=1}^{n_j} A_{ij}}{n_j}, \sqrt{\frac{\sum_{i=1}^{n_j} \left(A_{ij} - \left(\sum_{i=1}^{n_j} A_{ij} / n_j \right) \right)^2}{n_j}} \right) + \left(\frac{\sum_{i=1}^{n_j} B_{ij}}{n_j}, \sqrt{\frac{\sum_{i=1}^{n_j} \left(B_{ij} - \left(\sum_{i=1}^{n_j} B_{ij} / n_j \right) \right)^2}{n_j}} \right). \quad (2)$$

In I_k , k can be i, e, s , and a . I_i, I_e, I_s , and I_a represent illumination index, environmental index, scale index, and angle index, respectively;

In n_j , j can be 1, 2, 3, and 4. n_1, n_2, n_3 , and n_4 represent the number of different lights, environments, scales, and angles tested by the algorithm.

The scores of illumination adaptability, environment adaptability, scale adaptability, and angle adaptability can be calculated by Equation (3).

$$S_k = 3 * \frac{\sum_{i=1}^{n_j} A_{ij} / n_j}{\exp \left(\sqrt{\left(\sum_{i=1}^{n_j} \left(A_{ij} - \left(\sum_{i=1}^{n_j} A_{ij} / n_j \right) \right)^2 \right) / n_j} \right)} + \frac{\sum_{i=1}^{n_j} B_{ij} / n_j}{\exp \left(\sqrt{\left(\sum_{i=1}^{n_j} \left(B_{ij} - \left(\sum_{i=1}^{n_j} B_{ij} / n_j \right) \right)^2 \right) / n_j} \right)}. \quad (3)$$

In S_k , k can be i, e, s , and a . S_i, S_e, S_s , and S_a represent the light adaptability score, environmental adaptability score, scale adaptability score, and angle adaptability score, respectively.

3.3. Comprehensive Evaluation Index Formulation Based on Hierarchical Model. The comprehensive evaluation index model is constructed based on the hierarchical structure model.

3.3.1. Construction of Hierarchical Structure Model. We utilize analytic hierarchy process (AHP) to analyze the weight of the four indicators. The evaluation target layer, evaluation criteria layer, and evaluation scheme layer are used to build a three-layer structure model. The top-level evaluation target layer considers the purpose of decision-making. In the evaluation model, weights of illumination, environment, scale, and angle are obtained to get the final score.

The lowest evaluation scheme layer is an alternative scheme for decision-making, where seven test algorithms are selected, Faster RCNN [38], RetinaNet [39], ATSS [40], FoveaBox [41], GFocal Loss [42], PAFPN [43], and RepPoints [44], representing seven schemes. The middle layer is the criterion that needs to be considered in the evaluation process, namely, four evaluation indexes:

FIGURE 4: Quality score and speed score of the algorithm to be evaluated under different lighting, environment, angle, and scale conditions.

illumination index, environment index, scale index, and angle index. We have solved the relative weight problem of the evaluation criterion layer to the evaluation target layer. For each algorithm in the evaluation scheme layer, the final score is calculated according to the importance weight of the four indicators, so as to get the optimal scheme among the seven schemes.

3.3.2. Construct the Judgment Matrix. We construct the matrix using the consistent matrix method. Due to the high difficulty of comparison of factors with different properties, different understandings of the judgment of the four indicators in the evaluation criteria layer will lead to different grasp of the standards, resulting in evaluation errors. Through multidiscussion and expert consultation, a judgment table is constructed according to the scale table of paired comparison matrix. The scale table is as follows (see Table 1).

3.3.3. Hierarchy Single Sorting and Consistency Test. The judgment matrix contains subjective evaluation, and the weight of each index in the criterion layer is obtained according to the judgment matrix; then, a consistency test must be carried. If the verification is passed, the obtained weight can be used. Consistency index CI can be calculated as shown in Equation (4).

$$CI = \frac{\lambda_{\max} - n}{n - 1}. \quad (4)$$

We have to set a matrix dimension to find the average random consistency index RI. The RI value table is shown in Table 2.

The formula for calculating consistency ratio CR is as shown in Equation (5).

$$CR = \frac{CI}{RI}. \quad (5)$$

4. Result

4.1. The Subindex of Each Algorithm. On the algorithm test platform, we evaluate the performance of the algorithm model from several perspectives, such as illumination index, environment index, speed index, and angle index. For illumination index, light simulation and data sampling under

normal light, low light, and weak light are experimentally performed. As for the environmental index, the environmental simulation and data sampling under five scenes of constant light close range, thin fog, thick fog, rain, and snow are tested. For the angle index, three kinds of angle environment simulation and data sampling have experimented with angle 1, angle 2, and angle 3. For the scale index, the far and near environment simulation and data sampling of the distant, medium, and near scenes have experimented. The accuracy score and the different indexes' mean and variance value of the test algorithms for Faster RCNN, RetinaNet, ATSS, FoveaBox, GFocal Loss, PAFPN, and RepPoints are shown in Tables 3 and 4. The adaptability scores are shown in Table 5.

To more intuitively, Figure 5 shows the advantages of each algorithm in the four indicators.

It can be seen from the line chart that the angle adaptability score of the test algorithm is relatively high, which shows that all of the object detection algorithms have stronger strain ability in the practical task with various angles. On the other hand, it can be seen that the adaptability of the algorithms to the scene with various scales is slightly poor, so the scores are not high. In general, Foveabox has the best comprehensive performance, while RetinaNet performs the worst.

4.2. Comprehensive Evaluation Index. For different application scenarios, we set the judgment matrix differently. In some cases, climate conditions are different from those in ordinary areas due to variable meteorological conditions or air pollution and other factors. The conventional UAV remote sensing mission is susceptible to adverse and changeable weather. The application of some scenes is special, such as urban physical examination based on remote sensing, which limits factors such as UAV flight height and camera focal length. The targets in some scenes are densely arranged from multiple angles, and the UAV has a great limitation on the angle in the monitoring process.

4.2.1. Strong Illumination Index. Small and micro aerial vehicles usually fly under the clouds. If it is cloudy, the remote sensing pictures taken by the machine will not be blocked by the clouds but will be affected by the sharp changes in the illumination intensity, resulting in different

TABLE 1: Importance scale.

Contrast grade	Relatively important degree	Instructions
1	Same important	Both contribute equally to the goal
3	Slightly more important	From experience, one evaluation is slightly more favorable than the other
5	Fundamental importance	From experience, one evaluation is comparatively more favorable than the other
7	Really important	From experience, one evaluation is quite favorable than the other
9	Absolutely important	From experience, one evaluation is absolutely more favorable than the other
2, 4, 6, 8	The middle value of two adjacent degrees	Used when a compromise is needed

TABLE 2: RI value.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46	1.49	1.52	1.54	1.56	1.58	1.59

TABLE 3: mAP value of different algorithms on various scenes.

mAP	Run close	Slightly low light	Low light	Mist	Dense fog	Rain	Snow	Middle distance	Long distance	Oblique shot	Angle 1	Angle 2	Angle 3
Faster RCNN	0.555	0.553	0.597	0.677	0.470	0.568	0.602	0.082	0.000	0.369	0.784	0.485	0.917
RetinaNet	0.671	0.670	0.675	0.622	0.441	0.695	0.659	0.080	0.005	0.284	0.763	0.482	0.924
ATSS	0.499	0.447	0.379	0.460	0.275	0.429	0.424	0.040	0.000	0.299	0.755	0.558	0.817
FoveaBox	0.550	0.536	0.548	0.575	0.555	0.506	0.567	0.049	0.030	0.378	0.695	0.490	0.896
GFocal Loss	0.508	0.524	0.593	0.652	0.398	0.464	0.588	0.040	0.024	0.352	0.675	0.475	0.920
PAFPN	0.598	0.627	0.677	0.644	0.501	0.596	0.648	0.088	0.030	0.334	0.775	0.484	0.910
RepPoints	0.563	0.593	0.629	0.640	0.582	0.608	0.625	0.243	0.030	0.428	0.725	0.584	0.898

TABLE 4: Different indexes' mean and variance.

Index (mean, variance)	Environmental index	Angle index	Scale index	Light index
Faster RCNN	0.57, 4e-3	0.72, 0.03	0.21, 0.05	0.56, 4e-4
RetinaNet	0.61, 8e-3	0.72, 0.03	0.25, 0.08	0.67, 4e-6
ATSS	0.41, 5e-3	0.71, 0.01	0.17, 0.05	0.44, 2e-3
FoveaBox	0.55, 5e-4	0.69, 0.02	0.20, 0.05	0.54, 3e-5
GFocal Loss	0.52, 8e-3	0.69, 0.03	0.19, 0.05	0.54, 1e-3
PAFPN	0.59, 2e-3	0.72, 0.03	0.23, 0.06	0.64, 1e-3
RepPoints	0.60, 7e-4	0.73, 0.01	0.27, 0.04	0.59, 7e-4

TABLE 5: Adaptability score.

Adaptability score	Environmental	Angle	Scale	Light
Faster RCNN	3.681731	4.07472	2.577832	3.657885
RetinaNet	3.415586	3.69633	2.292505	3.610159
ATSS	3.462505	4.347446	2.723766	3.556002
FoveaBox	4.031839	4.411676	2.953402	4.002616
GFocal Loss	3.782423	4.243675	2.777054	3.853234
PAFPN	3.735508	4.065207	2.618862	3.887126
RepPoints	3.962834	4.332303	2.942333	3.932855

light and shade. Even in a sunny environment without the interference of clouds, the illumination is constantly changing. Although this change will not make a big difference in vision, it will have an impact on the gray value of the picture. In addition, the UAV is limited by the flight height and the focal length of the camera, so the coverage area of a single picture collected each time is small, and multiple images need to be stitched and fused into a panoramic image. The problem of local chromatic aberration caused by illumination deviation is also very difficult. For all-weather detection tasks in a changeable environment and application scenarios focusing on extracting geometric information of geographic space, illumination adaptability is in high demand.

FIGURE 5: Seven algorithm models' adaptability score in illumination, environment, angle, and scale.

4.2.2. Strong Environmental Index. The UAV has the characteristics of small size and lightweight, and it has higher requirements for takeoff and landing sites and is highly sensitive to the working environment, such as terrain, landform, and weather conditions. The performance of the same detection algorithm varies greatly in different weather conditions. If extreme conditions are encountered, the camera cannot shoot. For optical sensors, when clouds, fog, or water vapor interfere with the signal's path, the remote sensing image taken by the camera will be partially obscured. For the application of large changes in weather and lack of sunlight, it has requirements on environmental adaptability.

4.2.3. Strong Scale Index. Many of the objects in the scene are small and of varying sizes. Small scale objects show a smaller range of pixels in the image, while large scale objects show a larger range of pixels. Common remote sensing target monitoring is difficult to identify target objects at different scales with poor generalization ability. In the convolutional neural network, the deeper features have greater feelings and richer semantic information. However, if the geometrical information of the image is lost due to the reduction of the resolution, the shallow information has a small receptive field and rich geometrical information. For the application scenes with large change of target scale and target density, large image field of view, and complex background, adaptability to the environment is required.

4.2.4. Strong Angle Index. The objects in the scene generally rotate arbitrarily from multiple angles and present dense arrangement. The normal method of capturing the front frame is not enough. The single receptive field of neural networks does not have a good solution to the object variability. The receptive field of the general network is fixed in size and

has no angle, and it is only calculated along the horizontal direction. However, the vertical projection in the monitoring scene leads to different angles of the actual target. In moving object monitoring, target tracking, antidetection, and other application scenarios, the importance of angle index is relatively high. For tasks with large target rotation changes, it means that the targets of the same type have multiple angle changes, which requires higher angle adaptability.

4.2.5. The Military Object Detection Task. The multichannel style generation and conversion system of natural phenomena constructed in this project provide users with four seasonal changes of spring, summer, autumn, and winter and 24-hour day and night changes. Users can match any natural phenomena such as four seasons, day and night, rain and snow, and cloud and fog as required. The rendering of the scene in four seasons is shown in Figure 6.

The military scene is located in a mountainous area, and the environment is changeable, so the UAV detection task has a high adaptive demand to the environment. The shooting dimension is low, and the lighting changes are not drastic. Moreover, as the UAV focuses on detecting objects with multiple angle changes such as weapons and vehicles, the adaptive requirement for angle changes is also high. In the reconnaissance missions of small unmanned aerial vehicles, the flying height is relatively fixed, and the scale of the photos taken is not strict. For this task, the relative matrix X is selected after expert's review and questionnaire survey, as shown below, and the weight vector obtained by the geometric average method is $[0.14371197, 0.57484787, 0.0630452, 0.21839496]$. In the final comprehensive evaluation stage, for different object detection tasks, the evaluator can set the value of X by himself. However, it should be noted that this judgment matrix must be reviewed by a large

FIGURE 6: Spring, summer, autumn, and winter four seasons change renderings.

TABLE 6: The comprehensive evaluation scores of the seven algorithms.

Model	Faster RCNN	RetinaNet	ATSS	FoveaBox	GFocal Loss	PAFPN	RepPoints
Score	3.832837	3.54866	3.945057	4.175815	3.999654	3.887748	4.104338

number of questionnaires or the review of the expert in the field and must pass the consistency test.

$$X = \begin{bmatrix} 1 & 4 & 1/3 & 2 \\ 1/4 & 1 & 1/8 & 1/3 \\ 3 & 8 & 1 & 3 \\ 1/2 & 3 & 1/3 & 1 \end{bmatrix}. \quad (6)$$

We set n to be 4 and calculated the value of CI to be 0.017271 by the matrix X . Then, the value of CR was calculated as 0.019406. The criterion for whether the matrix passes the consistency test is to judge whether the CR value of the consistency ratio is less than 0.1. If the inequality is true, the consistency of the judgment matrix is determined to be acceptable; otherwise, the judgment matrix should be modified. The comprehensive evaluation scores of the seven algorithms involved in the evaluation are as follows (see Table 6).

5. Summary

We propose a novel method to solve the adaptability scores of various algorithms for different conditions in the UAV detection task and formulate a new benchmark. But the disadvantage is that the adaptability score of the algorithm in different environments depends on common comparisons with other algorithms. The more algorithms tested, the more accurate the results. Secondly, the main formula of the hierarchical model analysis method is based on qualitative analysis. In the judgment matrix, the importance assigned to each indicator is very arbitrary, which can cause conflicts in multiperson decisions. Although this article uses the expert decision-making method to summarize opinions and take weights, the process is long and cumbersome. To deal with the remote sensing target detection task with subtle environmental differences under a certain index, it is necessary to formulate the matrix again. The generalization ability of this method is not very brilliant. It cannot be flexibly applied to the complex remote sensing object detection scenes in the subdivided areas.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 61971383 and the CETC funding.

References

- [1] V. der Ouderaa, F. A. Tycho, and E. Daniel, "Reversible gans for memory-efficient image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4715–4723, California, 2019.
- [2] Z. Shen, M. Huang, J. Shi, X. Xue, and T. S. Huang, "Towards instance-level image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3678–3687, California, 2019.
- [3] Y. Alharbi, N. Smith, and P. Wonka, "Latent filter scaling for multimodal unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1458–1466, California, 2019.
- [4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, California, 2021.
- [5] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2242–2251, Venice, Italy, 2017.
- [6] H. Ray, H. Pfister, D. Silver, and T. A. Cook, "Ray casting architectures for volume visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 5, no. 3, pp. 210–223, 1999.
- [7] S. Mady and S. Abou El-Seoud, "An overview of volume rendering techniques for medical imaging," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, no. 6, pp. 95–106, 2020.

- [8] M. Levoy, "Display of surfaces from volume data," *IEEE Computer Graphics and Applications*, vol. 8, no. 3, pp. 29–37, 1988.
- [9] B. T. Phong, "Illumination for computer generated pictures," *Communications of the ACM*, vol. 18, no. 6, pp. 311–317, 1975.
- [10] J. T. Van Scheltinga, J. Smit, and M. Bosma, "Design of an On-Chip Reflectance Map," in *Proc. of the 10th EGWorkshop on Graphics Hardware*, pp. 51–55, Maastricht, The Netherlands, 1995.
- [11] P. Ljung, J. Krüger, E. Groller, M. Hadwiger, C. D. Hansen, and A. Ynnerman, "State of the art in transfer functions for direct volume rendering," *Computer Graphics Forum*, vol. 35, no. 3, pp. 669–691, 2016.
- [12] R. Krishna, Y. Zhu, O. Groth et al., "Visual genome: connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.
- [13] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1746–1754, Honolulu, Hawaii, 2017.
- [14] J. Zhang, W. Li, P. Wang, P. Ogunbona, S. Liu, and C. Tang, Eds., "A large scale RGB-D dataset for action recognition," in *International workshop on understanding human activities through 3D sensors*, Springer, Cham, 2016.
- [15] J. Straub, T. Whelan, L. Ma et al., "The replica dataset: a digital replica of indoor spaces," 2019, <http://arxiv.org/abs/1906.05797>.
- [16] T. Saaty, "Analytic Hierarchy Process," in *Encyclopedia of Bio-statistics*, Wiley, 2005.
- [17] J. Deng and W. Dong, "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, FL, USA, 2009.
- [18] T.-Y. Lin, M. Maire, and C. O. C. O. Microsoft, "Common objects in context," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 886–893, San Diego, 2005.
- [19] M. Everingham and J. Winn, "The PASCAL visual object classes challenge 2007 (VOC2007) development kit," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.
- [20] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: the KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [21] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.
- [22] W. Lin, H. Liu, S. Liu et al., "Human in events: a large-scale benchmark for human-centric video analysis in complex events," 2020, <http://arxiv.org/abs/2005.04490>.
- [23] S. Shao, Z. Zhao, B. Li et al., "CrowdHuman: a benchmark for detecting human in a crowd," 2018, <http://arxiv.org/abs/1805.00123>.
- [24] M. Cordts, M. Omran, S. Ramos et al., "The cityscapes dataset for semantic urban scene understanding," in *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, Las Vegas, Nevada, 2016.
- [25] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, 2016.
- [26] S. Qiao, W. Shen, W. Qiu, C. Liu, and A. Yuille, "ScaleNet: guiding object proposal generation in supermarkets and beyond," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.
- [27] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, 2017.
- [28] G. S. Xia, X. Bai, J. Ding et al., "DOTA: a large-scale dataset for object detection in aerial images," in *InCVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018.
- [29] K. Xu, X. Yang, B. Yin, and R. W. Lau, "Learning to restore low-light images via decomposition-and-enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2278–2287, Seattle, United States, 2020.
- [30] H. Long, Y. Chung, Z. Liu, and S. Bu, "Object detection in aerial images using feature fusion deep networks," *IEEE Access*, vol. 7, pp. 30980–30990, 2019.
- [31] C. Wang, X. Bai, S. Wang, J. Zhou, and P. Ren, "Multiscale visual attention networks for object detection in VHR remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 310–314, 2018.
- [32] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019, <http://arxiv.org/abs/1903.12261>.
- [33] C. Michaelis, B. Mitzkus, R. Geirhos et al., "Benchmarking robustness in object detection: autonomous driving when winter is coming," 2019, <http://arxiv.org/abs/1907.07484>.
- [34] E. Bruneton and F. Neyret, "Precomputed atmospheric scattering," *Computer Graphics Forum*, vol. 27, no. 4, pp. 1079–1086, 2008.
- [35] T. Aila and S. Laine, "PantaRay: fast ray-traced occlusion caching of massive scenes," *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, pp. 145–149, 2009.
- [36] M. Macklin and M. Matthias, "Position based fluids," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 1–12, 2013.
- [37] S. Green, "Screen space fluid rendering for games," in *Proceedings for the Game Developers Conference*, San Francisco, 2010.
- [38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [39] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, Venice, Italy, 2017.
- [40] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, United States, 2020.
- [41] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: beyond anchor-based object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 7389–7398, 2020.
- [42] X. Li, W. Wang, L. Wu et al., "Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection," 2020, <http://arxiv.org/abs/2006.04388>.

- [43] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, 2018.
- [44] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: point set representation for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea, 2019.