Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Camouflaged target detection based on multimodal image input pixel-level fusion[*]

Ruihui PENG[1,2], Jie LAI[†‡1], Xueting YANG[1], Dianxing SUN[1,3], Shuncheng TAN[3], Yingjuan SONG[1], Wei GUO[1]

[1]*Qingdao Innovation and Development Base, Harbin Engineering University, Qingdao 266000, China*

[2]*College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China*

[3]*Insitute of Information Fusion, Naval Aeronautical University, Yantai 264001, China*

[†]E-mail: laijie@hrbeu.edu.cn

Received July 26, 2023; Revision accepted Nov. 14, 2023; Crosschecked Aug. 28, 2024

Abstract: Camouflaged targets are a type of nonsalient target with high foreground and background fusion and minimal target feature information, making target recognition extremely difficult. Most detection algorithms for camouflaged targets use only the target's single-band information, resulting in low detection accuracy and a high missed detection rate. We present a multimodal image fusion camouflaged target detection technique (MIF-YOLOv5) in this paper. First, we provide a multimodal image input to achieve pixel-level fusion of the camouflaged target's optical and infrared images to improve the effective feature information of the camouflaged target. Second, a loss function is created, and the *K*-Means++ clustering technique is used to optimize the target anchor frame in the dataset to increase camouflage personnel detection accuracy and robustness. Finally, a comprehensive detection index of camouflaged targets is proposed to compare the overall effectiveness of various approaches. More crucially, we create a multispectral camouflage target dataset to test the suggested technique. Experimental results show that the proposed method has the best comprehensive detection performance, with a detection accuracy of 96.5%, a recognition probability of 92.5%, a parameter number increase of 1×10^4, a theoretical calculation amount increase of 0.03 GFLOPs, and a comprehensive detection index of 0.85. The advantage of this method in terms of detection accuracy is also apparent in performance comparisons with other target algorithms.

Key words: Camouflaged target detection; Pixel-level fusion; Anchor box optimization; Loss function; Multispectral dataset
https://doi.org/10.1631/FITEE.2300503 **CLC number:** TP391

1 Introduction

Camouflage is a self-defense mechanism formed during the long-term evolution of some organisms in nature (Lv et al., 2021). Chameleons and cuttlefish, for example, adjust to changes in their surroundings by changing their physiological characteristics to escape attacks from natural enemies. The goal of camouflaged target detection is to discover and identify targets concealed in the background environment (Tan et al., 2022). This type of target typically has a high degree of visual similarity to the backdrop environment in terms of color, brightness, and texture, and thus conventional reconnaissance methods can hardly detect it. Camouflage target detection has been widely considered by various military powers as a detection technology under a complex backdrop, and it is regarded as the key to improving battlefield situation information. In addition to having high military research value, research on camouflage target detection is helpful in promoting

[‡] Corresponding author

[*] Project supported by the Shandong Provincial Natural Science Foundation of China (No. ZR2020MF015) and the Aerospace Science and Technology Innovation Institute Stabilization Support Project (No. ZY0110020009)

ORCID: Jie LAI, https://orcid.org/0009-0005-7918-0941

the rapid development of pest control in agriculture (Fan et al., 2020a) and polyp detection in medicine (Fan et al., 2020b).

Optical camouflage and infrared (IR) camouflage are two methods commonly used in military applications to improve the fusion degree of the target and the surrounding background and reduce the target's detectability. The former alters the target's apparent characteristics such as color, texture, and brightness (Zhang et al., 2022), whereas the latter alters its emissivity or surface radiation temperature. Some camouflage effects are so intense that the human eye cannot distinguish them (Liang et al., 2021), posing significant challenges to target positioning and recognition on the battlefield. Furthermore, due to their uniqueness, military targets frequently appear in complex field environments such as jungles, deserts, and snow, further increasing the difficulty of detection. Combined with the preceding research and analysis, the difficulties of military camouflage target detection primarily lie in the two points below: (1) Camouflage means diversification. In today's battlefield, conventional single camouflage has difficulty in meeting the needs of practical applications. Military targets must counter not only visible light band reconnaissance, but also IR and other band threats. (2) Single-modal images have a limited amount of feature information. The visible light band camouflaged target has rich visual features such as brightness and texture, but the position information is blurred. The camouflaged target's position in the IR band is obvious, but textural information is scarce.

Traditional camouflaged target detection algorithms involve mainly histograms of oriented gradients (HOGs), support vector machines (SVMs), deformable part models (DPMs), and other methods. Wu et al. (2015) used three-dimensional (3D) convex surface detection operators to solve the problem of the edge detection algorithm not being able to detect camouflaged targets with 3D convex surfaces. Bhajantri and Nagabhushan (2006) proposed a technique for detecting targets hidden in the environment. First, the texture feature algorithm based on the co-occurrence matrix was used to calculate a small area of the image. Second, the clustering analysis algorithm was used to detect the defect part in the image. Finally, the watershed segmentation algorithm was used to locate the camouflaged target. The similarities between the flaw and background textures were used to identify the camouflaged target. To summarize, this type of algorithm has a limited ability to extract features, resulting in the utilization rate of target feature information being low and the effective feature information being seriously lost, making it unsuitable for use in the complex field environment of a battlefield.

In 2012, the detection accuracy of the image identification challenge with Krizhevsky's AIexNet was 10.9% higher than that with the conventional target recognition approach, indicating that target detection technology has entered the deep learning era. At present, deep learning based object detection algorithms can be simply divided into two categories: one-stage detection model and two-stage detection model. The two-stage method first generates a series of candidate boxes as samples and then classifies the samples using convolutional neural networks (CNNs). Representative algorithms include Region-CNN (R-CNN) (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), and Faster R-CNN. The one-stage method does not need to generate candidate boxes and directly turns the problem of target box positioning into a regression problem, which greatly accelerates the network prediction. Representative algorithms include the single shot multibox detector (SSD) (Liu W et al., 2016), RetinaNet (Lin et al., 2020), and the YOLO (you only look once) series (Redmon et al., 2016; Redmon and Farhadi, 2017; Bochkovskiy et al., 2020).

At present, the target detection algorithm based on deep learning has been widely used in military fields such as precision guidance and battlefield reconnaissance. Zhang et al. (2022), for example, proposed an MC-YOLOv5s detection algorithm for camouflage personnel detection based on camouflaged target characteristics. Cheng Y et al. (2022) proposed an attention-based neighbor selective aggregation network (ANSA-Net) that can effectively detect camouflaged objects. Yadav et al. (2018) proposed a target detection model that successfully distinguishes camouflaged targets by combining hyperspectral and LiDAR data at the decision level. Hu et al. (2015) used wavelet transform to incorporate near-IR band image information into the R, G, and B channels of visible light, which improved the details of visible light images while supplementing the target's near-IR information and had more advantages in terms of target recognition and camouflage evaluation. Putatunda et al.

(2022) proposed a multimodal head object detection algorithm that combines camouflaged targets' visible image features and depth image features to compensate for the shortcomings of insufficient visible image positioning information. In summary, the camouflaged target detection algorithm based on deep learning has a better ability to extract and integrate target feature information. It can not only effectively fuse information from different data sources but also fully exploit data advantages to improve the performance of camouflaged target detection, resulting in good applicability.

In this paper, based on the characteristics of the YOLOv5 algorithm, a multimodal image input port is constructed at the input end of the network. The pixel-level fusion strategy is used to realize the full utilization of the feature information of the camouflaged target optical (RGB) image and the IR image. The training process of the model is optimized by the loss function and the anchor frame optimization algorithm to improve the robustness of the algorithm, and the problems of low detection accuracy and high missed detection rate of camouflaged targets are solved. Our main contributions are as follows:

1. Multispectral data collection is created which includes RGB and IR images of camouflaged targets in the same state space.

2. An end-to-end multimodal image fusion technique is proposed, and a military camouflage target evaluation system is built.

2 Algorithm improvement

2.1 Introduction to the YOLOv5 algorithm

The YOLOv5 algorithm is a one-stage detection algorithm that is both accurate and quick. As shown in Fig. 1, the algorithm is divided into four parts: Input,

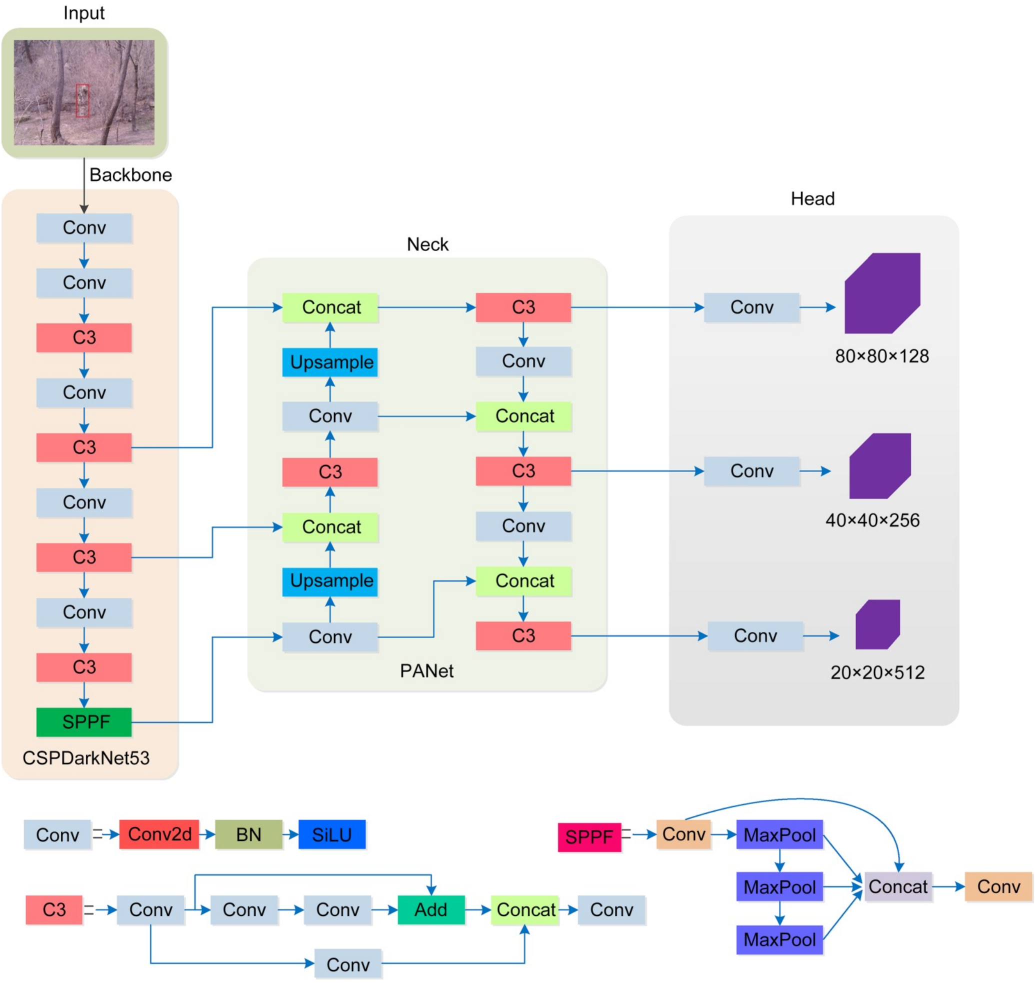

Fig. 1 Network structure diagram for the YOLOv5 algorithm

Backbone, Neck, and Head. The first is the input terminal. Input can meet the three-channel image loading. The input image is scaled to the network's input size (640×640) in the image preprocessing stage, and the mosaic data enhancement is used to improve the model's training speed and network accuracy in the network training stage. Backbone uses mainly CSP-DarkNet53 as the reference network to extract some common feature representations of the input image, including components such as Conv, C3, and spatial pyramid pooling - fast (SPPF). The Conv module is composed of convolution, a batch normalization (BN) layer, and an activation function (SiLU) to realize the conversion and extraction of input features. The C3 module is composed of a Conv module, connection (Concat), and superposition (Add), aiming to improve the feature extraction ability by increasing the depth and receptive field of the network. The SPPF module is composed of the Conv module, MaxPool, and Concat, aiming to realize the spatial invariance and position invariance of input data. Neck uses the PANet module to achieve different levels of feature fusion. Specifically, first the top-down feature pyramid is used to achieve upsampling and fusion with coarse-grained feature maps to achieve the fusion of different levels of features. Finally, the bottom-up feature pyramid is used to fuse different levels of feature maps. The Head module outputs three feature maps, which are used to detect small, medium, and large targets of different scales. The feature map sizes (length×width×number of channels) are 20×20×512, 40×40×256, and 80×80×128.

2.2 Pixel-level fusion

The imaging of camouflaged targets varies significantly across bands, such as the most common RGB and IR images. The optical sensor captures the reflected light and generates the RGB image (Liu CX, 2022). It has many details, such as brightness, color, edge, and texture. The camouflaged target weakens its own exposed features (such as brightness and texture) through active or passive methods, achieving better integration with the surrounding complex environment, reducing the target's saliency in the background and making it difficult to find. In addition, visible light sensors are susceptible to external environmental factors such as visibility, light, and smoke (Qi, 2022), resulting in poor optical image quality and missing effective target information. The IR image is obtained by the IR sensor capturing the energy difference between the target and the scene, which is related to the surface temperature of the target and its own emissivity. As a consequence, the IR sensor has strong anti-interference capabilities and can continue to function normally in harsh environments, such as smoke, snow, and at night. Therefore, longwave IR (LWIR) equipment is a common method of reconnaissance on the battlefield. In contrast to the visible image's three-channel R, G, B display, the original IR image is recorded in a single-channel gray image matrix with a value ranging from 0 to 255. Because the IR image reflects thermal radiation information in the detection scene (Sun et al., 2023), the larger the difference in temperature between the target and the background is, the greater the saliency of the target in the image is. However, because its imaging band is invisible, it is incompatible with human vision and lacks environmental information such as image detail texture and illumination.

RGB images contain rich texture detail information and are more in line with human visual perception for camouflage target recognition, but the target location information in the image is blurred, making it harder to fully exploit the target detection algorithm's potential. In a complicated background environment, IR images can highlight the target's position information, effectively compensating for the drawbacks of simplify using RGB images in the target detection process. As a result, fusing RGB and IR images in the same state space, taking full advantage of the complementary characteristics of the two, and fully exploiting the advantages of multimodal data information, can not only compensate for the shortcomings of insufficient visible light image position information of camouflaged targets and blurred edge contours of IR images, but also improve the detectability of targets and scenes in camouflaged environments.

Multimodal image fusion is the technique of combining numerous images with multiple features into a single image. It seeks to integrate the complimentary information in the source image and provide a fused image with rich texture details that highlight the main object. The pixel-level fusion of images is the most basic and fundamental image fusion method. A specific rule is employed at the pixel level to fuse a multisource image of the target in the same state space into a fused image with more detailed texture and edge contours. Following the completion of the fusion task, follow-up operations, such as feature extraction and target recognition, can be performed based on the real demands. Because pixel-level fusion processes the pixel data directly, the pixel data information can be completely exploited in the fusion process without significant loss, and the original detailed information of the scene target can be kept to the possible greatest extent. There are four types of pixel-level image fusion methods: spatial domain methods that directly process pixel values, multiscale and multiresolution based transform domain methods, fusion methods based on mathematical theoretical models, and hybrid fusion methods that combine multiple methods. The methods described above work on the original data in the transform or spatial domain. When used for target identification, challenges such as complex processing flow and low ablation efficiency severely limit the practicability of pixel-level image fusion in target detection technology.

As a result, in this study we propose a multimodal image input port that allows for the simultaneous loading of RGB and IR images, allowing for pixel-level fusion of these two images. On this basis, feature extraction, target identification, and other links are carried out directly, which not only increases the effect of detecting camouflaged targets but also improves the practicability of pixel-level image fusion in target detection. The basic idea behind this method is to improve the loading capability of the YOLOv5 algorithm's input to multichannel data, allowing for the integration of two different modes of information of a three-channel RGB image and a single-channel IR image. First, the number of convolution input channels in Backbone's first layer is increased to four, allowing for the simultaneous loading of the three-channel RGB image and single-channel IR image data. Second, the approaches to augmentation of mosaic data and output of network prediction results are tweaked to improve the model's generalization capacity and visualize the output results. During network training, RGB and IR image pairs in the same state space are simultaneously loaded into the network's input port, and the feature extraction and target recognition of RGB and IR images are performed via a CNN. No information is lost during later data processing because RGB and IR information is immediately entered into the network. Fig. 2 depicts the improved network architecture.

2.3 Anchor box optimization

The anchor box is a predefined bounding box with a specific height and breadth that represents the prospective target's initial state and is an effective way to obtain the target's potential distribution area. The YOLOv5 method is an anchor box detector. The input image is divided into $m\times n$ regions during algorithm execution, and then a number of anchor boxes are generated at the middle of each grid according to the preset aspect ratio as candidate regions. The model forecasts whether these candidate regions contain objects and, if so, what type of objects they contain. The YOLOv5 algorithm's predefined anchor frame clusters the real frames of all targets in the COCO dataset to generate nine clustering centers. The camouflaged targets studied in this research are mostly camouflaged soldiers, and most targets are tall with respect to their other dimensions.The original anchor frame cannot now meet the needs of camouflaged personnel target detection. The *K*-Means algorithm is a straightforward, practical, and easy-to-implement unsupervised clustering method. The *K*-Means algorithm is used to calculate the original anchor boxes of YOLOv5. However, it must first randomly select *K* cluster centers before clustering; the final clustering result is highly dependent on cluster center initialization.

In this study, the *K*-Means++ algorithm is used to replace the original clustering algorithm. This algorithm optimizes the random selection center of the first step in the *K*-Means algorithm, so that the randomly selected center point no longer tends to the local optimal solution but instead is allowed to be as close to the regional global optimal solution as possible, thus improving the detection accuracy and effect to some extent. Furthermore, the original *K*-Means clustering algorithm uses the Euclidean distance to calculate the distance between the sample and the cluster center, but this method produces a larger error on the large anchor box than on the small anchor box, and the purpose of clustering is that there is a larger intersection over union (IoU) between the

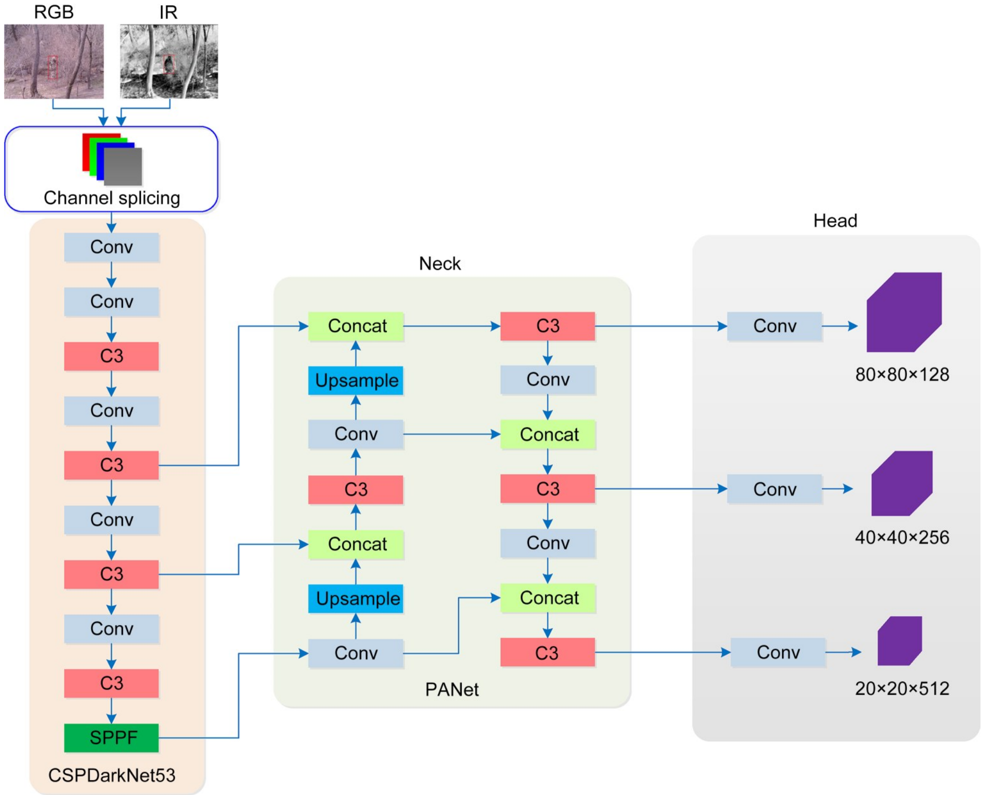

Fig. 2 Block diagram of the pixel-level fusion method based on the 4-channel port

anchor box and the adjacent prediction box, which is unrelated to the actual size of the anchor box. Accordingly, IoU is used as the metric evaluation standard of the *K*-Means++ clustering algorithm in this study. The IoU value increases as the sample gets closer to the cluster center. The IoU value must be negatively correlated to meet the definition of the loss function:

$$\mathrm{IoU} = \frac{A \cap B}{A \cup B}, \tag{1}$$

$$D(A,B) = 1 - \mathrm{IoU}(A,B), \tag{2}$$

where *A* and *B* represent the anchor box and the ground-truth box, respectively.

All of the ground-truth boxes in the dataset adopted in this study are optimized using the improved clustering algorithm, and the results are shown in Table 1. The optimization results improve the degree of matching between the anchor box and the target prediction box even more. A better initialization is given to the model at the start of network training, which accelerates the convergence of the network model and ultimately improves the detection effect.

Table 1 Optimized anchor box using the *K*-Means++ algorithm

Feature map scale	Anchor box size		
	Box 1	Box 2	Box 3
Big	(10, 25)	(31, 35)	(65, 48)
Middle	(32, 96)	(45, 89)	(44, 127)
Small	(57, 134)	(59, 177)	(79, 216)

2.4 New loss function

The loss function is used to assess the degree of inconsistency between the model's predicted and true values to assess the quality of model prediction and to play a penalty role in the target detection task. It should be kept to a minimum during training. The smaller the loss function is, the closer the model's predicted value is to the true value. YOLOv5's loss function is divided into three parts: confidence loss function (L_{conf}), category loss function (L_{cls}), and position loss function (L_{CIoU}). Among them, the confidence loss function and the category loss function are described by the binary cross entropy function, and the position loss function is represented by CIoU. The CIoU loss function is calculated by aggregating bounding box regression indicators such as the distance between the prediction boxes and ground-truth boxes, the overlapping area, and the aspect ratio. However, it does not account for the mismatch direction between the prediction boxes and the ground-truth boxes, resulting in prediction box position drift during the model training process, which reduces not only the model's convergence speed but also its efficiency. To that end, we employ a new structured intersection over union (SIoU) (Gevorgyan, 2022) position loss function, as shown in Eqs. (3)–(7). The loss function has directionality based on the original bounding boxes' regression index by measuring the vector angle between the prediction box and the regression box, which can cause the prediction box to move quickly to the nearest axis, which is beneficial to the convergence process and training effect.

$$L_{\text{box}} = 1 - \text{IoU} + \frac{\varDelta + \varOmega}{2}, \tag{3}$$

$$\varLambda = 1 - 2\sin^2(\arcsin x - \pi/4), \tag{4}$$

$$\gamma = 2 - \varLambda, \tag{5}$$

$$\varDelta = \sum\nolimits_{t=x,y}(1 - \mathrm{e}^{-\gamma\rho_t}), \tag{6}$$

$$\varOmega = \sum_{t=w,h}(1 - \mathrm{e}^{-\omega_t})\theta, \tag{7}$$

where $\varDelta$ denotes the distance loss, $\varOmega$ is the shape loss, θ represents the coefficient of shape loss, $\varLambda$ represents the angular loss of the prediction box moving quickly to the horizontal or vertical lines of the ground-truth box, and w and h represent the width and height of the ground-truth box's and prediction box's minimum bounding rectangles, respectively. Furthermore, because the dataset scene used in this study has only foreground (camouflage personnel) and backdrop, there are no multiclass detection requirements, so the class loss function can be removed. Based on the research presented above, we suggest a new loss function, as illustrated in Eqs. (8)–(10):

$$\text{Loss} = L_{\text{conf}} + L_{\text{SIoU}}, \tag{8}$$

$$\begin{aligned} L_{\text{conf}} = {} & \lambda_{\text{obj}}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{\text{obj}}\,[-\hat{C}_i \ln C_i - (1-\hat{C}_i)\ln(1-C_i)] \\ & + \lambda_{\text{nobj}}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{\text{nobj}}\,[-\hat{C}_i \ln C_i - (1-\hat{C}_i)\ln(1-C_i)], \end{aligned} \tag{9}$$

$$L_{\text{SIoU}} = \sum_{i=0}^{S^2}\sum_{j=0}^{B} L_{\text{box}}. \tag{10}$$

Herein S signifies the mesh size. B denotes the number of prediction boxes per grid. I_{ij}^{obj} indicates whether the j^{th} prediction box of the i^{th} grid has a predicted target—if yes, the number is 1; otherwise, the number is 0. I_{ij}^{nobj} indicates whether the i^{th} grid's j^{th} prediction box has a target that does not need to be predicted—if so, the number is 1; otherwise, the number is 0. λ_{obj} represents the prediction box's goal balance coefficient. λ_{nobj} represents the prediction box's non-target balance coefficient. C_i is the prediction box's confidence. $\hat{C}_i$ is the actual box's confidence.

3 Multispectral dataset

There is currently no public dataset available to support research into camouflaged target detection using multimodal fusion technology. To address this issue, the research team began collecting data at an early stage. The proposed multispectral dataset includes two subsets of RGB images and IR images, reflecting the imaging information of the camouflaged target in the visible and LWIR bands. The target type in this dataset is camouflage personnel with different postures (such as creeping, standing, and half squatting) and different backgrounds. It contains 2000 pairs of RGB images and IR images. Each pair of images was collected in the same time and space, and the registration accuracy of the two is high, which can meet the pixel-level fusion requirements. Red, green, and blue make up the three color channels in an RGB image, but there is only one gray channel in an IR image. The occupied space is 1.2 MB for the IR image and 487 KB for the RGB image, both of which are 1024×768 in size. During the model training phase, the camouflaged target dataset was separated into three parts: training (1500), verification (400), and testing (100). The training and verification sets were drawn at random from the same dataset and had significant data similarity. To ensure the model's fairness during the testing process, the data used in the testing set differed from the data used in the training and verification sets. Labeling software was then used to label the targets in the training and verification sets.

The experimental personnel wore woodland camouflage clothing and desert camouflage clothing with optical and IR camouflage effects while collecting data. To simulate the real battlefield environment, the natural environmental background was used to hide the experimental personnel in woodlands, deserts, valleys, lawns, and other environments. The data acquisition equipment used was a high-resolution thermal imaging camera model T1050sc from FLIR, capable of acquiring both LWIR images and visible light images with a pixel size of 1024×768. In addition, typical time periods, such as morning, noon, and evening, were selected for data acquisition, and the detection distance was about 100–200 m. The data acquisition personnel’s hand-held dual-band IR thermal imager took images of camouflaged targets in both visible and IR bands from varying heights, angles, and distances. The quality of visible light images is closely linked to the present ambient light intensity; for IR images, the background temperature difference, in addition to its own emissivity, is a significant factor. In general, optical camouflage is more effective under poor lighting conditions and dense vegetation, and IR camouflage is more effective when the ambient temperature difference is small. As a result, during the data acquisition process, we should consider not only the change in environmental illumination but also the effect of temperature change on the final imaging data to enrich the dataset type, improve data quality, and strengthen the model’s generalization ability for different environments. Fig. 3 depicts some of the dataset’s objectives. We proposed a multispectral dataset with a high degree of fusion between the foreground and background, and the target’s exposure characteristics are fewer. It is akin to a real-world battlefield and has high research value.

4 Experiments

The research group’s multimodal image camouflaged target dataset was used to validate the suggested algorithm. The following is the primary environment configuration of the experimental platform: AMD R7 6800H CPU, 16 GB RAM, Nvidia GeForce RTX3060 6 GB GPU, Win11 professional operating system, Python3.9.12 language, CUDA 11.3 parallel computing design, CUDNN 7.6.5 deep neural network acceleration library, and the PyTorch1.1.0 deep learning framework.

4.1 Evaluation metrics

The accuracy and speed of detection can be used to quantify the quality of the target detection model. Precision (P), Recall Rate (R), Average Precision (AP), and mean Average Precision (mAP) are the evaluation metrics. The number of frames detected per second (FPS) is used to calculate the speed evaluation index. In this study, P, R, mAP, and FPS are used to preliminarily

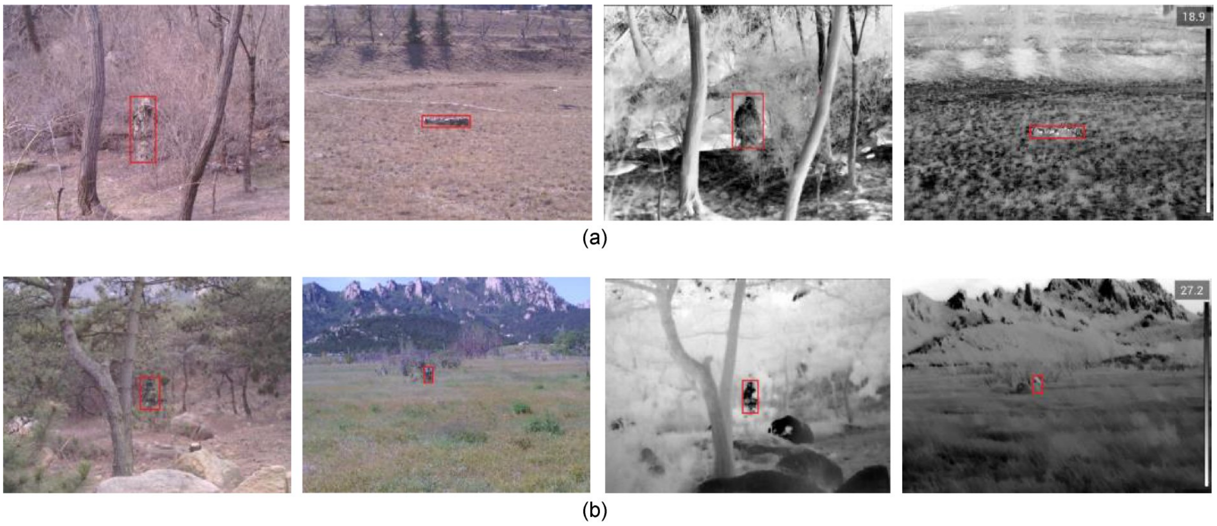

Fig. 3 Dataset of multimodal image camouflaged targets: (a) desert setting; (b) woodland setting
Left two, visible camouflaged targets; right two, infrared camouflaged targets

evaluate the target detection model. The calculation formulae for P, R, AP, and mAP are shown in Eqs. (11)–(14). FPS is obtained by calculating the average time of detecting 100 images in the testing set on the Nvidia GeForce RTX3060 hardware device. Furthermore, N_{para} represents the number of parameters in the model, and the computation formula is provided in Eq. (15). The algorithm complexity is measured using the GFLOPs (Giga floating-point operations) metric, which represents one billion floating-point operations executed per second, and the calculation formula is given as Eq. (16).

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{11}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{12}$$

$$\text{AP} = \int_0^1 P(r)\,\mathrm{d}r, \tag{13}$$

$$\text{mAP} = \frac{1}{m}\sum_{i=1}^{m}\text{AP}_i, \tag{14}$$

$$N_{\text{para}} = \{[(K_{\text{w}}K_{\text{h}})C_{\text{in}}]C_{\text{out}}\}C_{\text{out}}, \tag{15}$$

$$N_{\text{GFLOPs}} = 2C_{\text{out}}H_{\text{out}}W_{\text{out}}(C_{\text{in}}K_{\text{w}}^2 + \text{bias}). \tag{16}$$

As shown in Table 2, the above formulae represent four possible target detection results: TP (true positive), FP (false positive), TN (true negative), and FN (false negative); K_{w} represents the width of the convolution kernel; K_{h} represents the height of the convolution kernel; C_{in} represents the convolution kernel's input channel; C_{out} represents the convolution kernel's output channel; H_{out} and W_{out} represent the height and width of the input feature map, respectively; bias represents the bias term.

Table 2 Test result classification and interpretation

Metric	Implication
TP	Positive class prediction as positive class
FP	Negative class prediction as positive class
TN	Negative class prediction as negative class
FN	Positive class prediction as negative class

Given the uniqueness of the military context, evaluating the performance of a military camouflage target recognition algorithm solely on the mAP and FPS of the target detection model is insufficient to satisfy the practical application requirements. The algorithm's actual performance must be evaluated in the context of the related military application. The indexes of missed alarm (MA) and false alarm (FA) for military camouflage target detection are created after comparing the concepts of missed alarm and false alarm in radar target detection.

$$\text{MA} = \sum_{i=1}^{n}(N_i - D_i)\Big/\sum_{i=1}^{n}N_i, \tag{17}$$

$$\text{FA} = \sum_{i=1}^{n}E_i\Big/\sum_{i=1}^{n}N_i, \tag{18}$$

where N_i represents the number of camouflaged targets in the i^{th} image, D_i the number of camouflaged targets correctly predicted by the detection algorithm in the i^{th} image, E_i the number of non-camouflaged targets recognized as camouflaged targets in the i^{th} image, and n the total number of images participating in the detection.

If each evaluation index is evaluated independently in the actual evaluation of the performance of the camouflaged target detection algorithm, the conclusions obtained are often isolated, making it difficult to evaluate the performance from the perspective of the whole and the system, and the comprehensive performance of the algorithm cannot be quantitatively described. Some of the above indicators must be positively correlated and normalized to unify the dimension and better measure the performance of the target detection algorithm under different indicators. MA and FA are two of the evaluation indexes, with a value range of [0, 1]. The lower the value is, the better the camouflaged target detection algorithm performs. Therefore, it is necessary to perform positive correlation processing. As shown in Eqs. (19) and (20), $\overline{\text{MA}}$ represents the missed alarm rate after positive correlation processing, and $\overline{\text{FA}}$ represents the false alarm rate after positive correlation processing.

$$\overline{\text{MA}} = 1 - \text{MA}, \tag{19}$$

$$\overline{\text{FA}} = 1 - \text{FA}. \tag{20}$$

Furthermore, FPS must be linearly normalized. Given the large differences in target detection speed between algorithms, the obtained FPS values have a high degree of differentiation. To limit the original discrete data to [0, 1], the logarithmic nonlinear normalization method is used. As shown in Eq. (21), this method can avoid the impact of extreme values on the final evaluation results.

$$\overline{F}=\frac{\lg F}{\lg\max F},\tag{21}$$

where $\overline{F}$ represents the normalized FPS in the above formula, and $\max F$ reflects the maximum value in FPS. Following the aforementioned processing, all indicators are positively correlated with the performance of the camouflaged target detection algorithm, with a value range of [0, 1]. The comprehensive detection capability index C of the camouflaged target is suggested from the comprehensive assessment of the camouflaged target detection performance based on the above mAP, $\overline{F}$, $\overline{\text{MA}}$, $\overline{\text{FA}}$, and other indicators. Eqs. (22) and (23) provide a detailed explanation of this index.

$$C=\begin{bmatrix}W_1 & W_2 & \cdots & W_\alpha\end{bmatrix}\begin{bmatrix}\lambda_1\\ \lambda_2\\ \vdots\\ \lambda_\alpha\end{bmatrix},\tag{22}$$

$$\sum_{\alpha=1}^{N}W_\alpha=1,\tag{23}$$

where W_α is the weight size, which can adjust the attention to various performance indicators in camouflage target detection based on the weight of different indicators, and the indicator has good openness and can constantly expand the subindicators based on real requirements. Its value range is [0, 1] and it fulfills Eq. (23). λ_α represents the index subitem, including mAP, $\overline{F}$, $\overline{\text{MA}}$, $\overline{\text{FA}}$, and other indicators, while N represents the total number of subindicators.

4.2 Parameter values for training

The official YOLOv5 pretraining weight was chosen during the experiment, and the model parameters can be properly initialized at the start of the training. During training, the optimizer chose stochastic gradient descent (SGD) and adjusted the input image size to 640×640. The initial learning rate was set to 0.005 and progressively decreased over time to 0.0005. The momentum component was set to 0.937, and the weight attenuation coefficient was set to 0.0005. The number of loaded images in each batch was set to 8. To minimize overfitting, we employed the YOLOv5 series' minimum model, which had just 7.2×10^6 parameters, the fewest layers, and the lowest computing complexity, with 50 training cycles.

4.3 Ablation experiments

4.3.1 Experimental process

We conducted ablation research, set up eight sets of experiments, and verified them on the camouflage target dataset to verify the efficacy of each method in the multimodal image fusion YOLOv5 (MIF-YOLOv5) algorithm. Furthermore, to ensure fairness of the experiments, the same parameters were specified for each variable. The confidence threshold and non-maximum suppression IoU threshold were set to 0.5 and 0.01, respectively, during the production of the test results. The mAP index curves of various methods on the camouflaged target dataset are shown in Fig. 4. In the training procedure, Fig. 5 depicts the loss function curves of the training and validation sets. Table 3 displays the performance indicators for various methods on the dataset.

a: The performance of the YOLOv5 algorithm was verified on the RGB dataset;

b: The performance of the YOLOv5 algorithm was verified on the IR dataset;

c: Verification was carried out on the YOLOv5 algorithm optimized by the anchor box on the IR dataset;

d: The YOLOv5 algorithm was validated on the IR dataset using a new loss function;

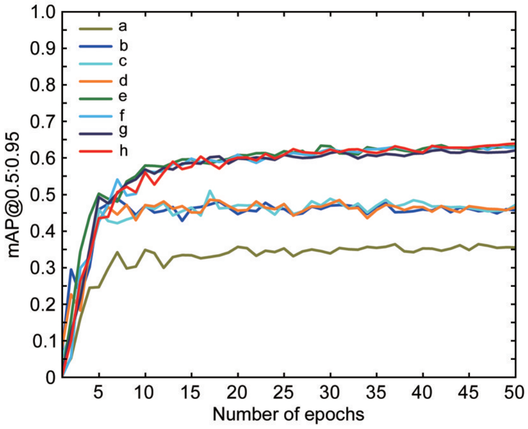

Fig. 4 The mAP@0.5:0.95 index curves of different experimental groups (References to color refer to the online version of this figure)

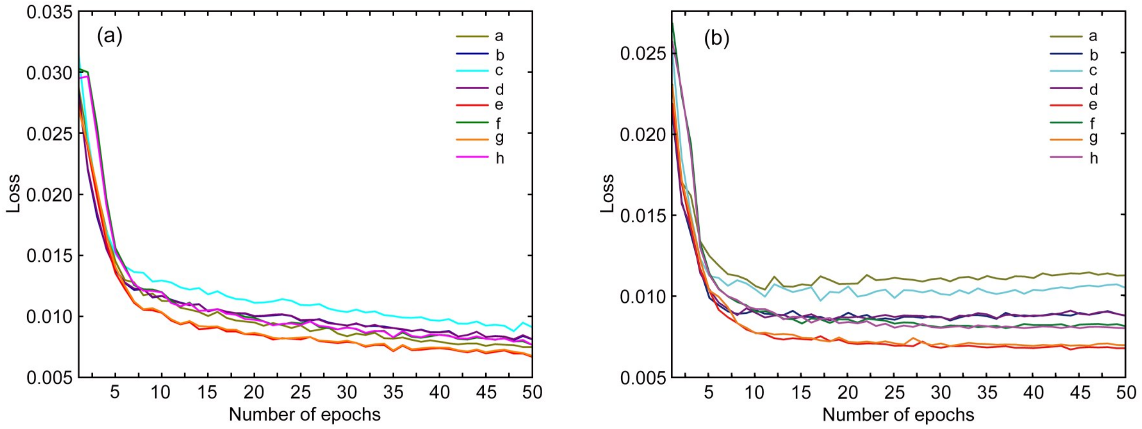

Fig. 5 The loss function change curves on the training set (a) and verification set (b) (References to color refer to the online version of this figure)

Table 3 Performance comparison of different experimental groups

Class	Number of parameters ($\times10^6$)	GFLOPs	P	R	mAP@0.5:0.95	Frame rate (frame/s)	MA	FA
a	6.7	15.9	0.894	0.813	0.364	78	0.149	0.037
b	6.7	15.9	0.896	0.817	0.488	76	0.205	0.009
c	6.7	15.9	0.906	0.847	0.511	77	0.187	0.009
d	6.7	15.9	0.926	0.854	0.485	79	0.177	0.019
e	7.0	16.2	0.977	0.964	0.635	71	0.131	0.018
f	7.0	16.2	0.968	0.964	0.633	70	0.093	0
g	7.0	16.2	0.968	0.967	0.622	71	0.084	0.009
h	7.0	16.2	0.965	0.966	0.639	67	0.075	0.019

e: Pixel-level fusion of RGB and IR data was realized using a four-channel input port;

f: The optimized anchor frame was used based on experimental group e;

g: The new loss function was used based on experimental group e;

h: The optimized anchor frame and the new loss function were used based on experimental group e.

4.3.2 Analysis of experimental findings

For experimental groups a and b, the performances of the original YOLOv5 algorithm were assessed on the camouflage target's RGB and IR datasets, respectively, serving as the benchmark for subsequent method comparison. The detection performance of YOLOv5 on the IR dataset was better than that on the visible dataset, as shown in Fig. 4 and Table 3. The mAP@0.5:0.95 of experimental group b (the average mAP at various IoU levels was represented) increased by 0.124, FPS decreased by 2 frames/s, MA increased by 0.056, and FA decreased by 0.028 when compared to experimental group a. This demonstrated that the IR band characteristics of camouflaged targets are significant and easy to detect. Experimental group c was based on experimental group b and it re-optimized the anchor box using the *K*-Means++ clustering algorithm. When compared to experimental group b, its mAP@0.5:0.95 increased by 0.023, MA decreased by 0.018, *P* increased by 0.01, and FPS had the minimum influence. This demonstrated that the clustering algorithm can improve the matching degree between the anchor box and the ground-truth box to some extent, allowing the model to perform better at the start of training. For experimental group d, the new loss function was used to evaluate the difference between the predicted value and the real value of the model based on experimental group b. When compared to experimental group b, *P* increased by 0.03, *R* increased by 0.037, mAP@0.5:0.95 decreased somewhat, MA decreased by 0.028, and FPS and FA increased marginally, indicating that the new loss function improved model convergence speed and reduced the missed detection rate while improving detection accuracy. The results of experimental group e showed that, following pixel-level fusion of RGB and IR images, the number of parameters of the network model increased by 106 556, and the GFLOPs increased by 0.3. At the same time, when compared to experimental groups a and b, P grew by 0.083 and 0.081, R increased by 0.151 and 0.147, and mAP@0.5: 0.95 increased by 0.271 and 0.147, respectively. The change in the above index values indicated that the fusion of visible light and IR image information of camouflaged targets can compensate for the lack of single-modal image features, but it also increased data processing and noise, resulting in a decrease in FPS and an increase in MA. In comparison to experimental group e, P, R, mAP@0.5:0.95, and FPS did not vary significantly in experimental group f; however, FA decreased by 0.018 and MA decreased by 0.038. The mAP@0.5:0.95 of experimental group g was somewhat reduced compared to that of experimental group e, while the FPS remained unchanged. The changes in P, R, FA, and mAP@0.5: 0.95 in experimental group h were smaller than those in experimental group e, but the missed detection rate was reduced by only 0.056.

From the standpoint of military applications, the contributions of the four evaluation indexes proposed in this research to the performance evaluation of camouflage target detection algorithms from large to small were MA, mAP@0.5: 0.95, FPS, and FA, and the weights were 0.4, 0.3, 0.2, and 0.1, respectively. The MA and FA were positively correlated using Eqs. (19) and (20), and the FPS was normalized using Eq. (21). The comprehensive detection ability index of camouflaged targets of different methods was obtained by Eq. (22), and the results are shown in Table 4.

Table 4 Comprehensive performance comparison among different experimental groups

Group	C	Group	C
a	0.74	e	0.83
b	0.73	f	0.85
c	0.75	g	0.83
d	0.79	h	0.85

Table 4 shows that the three methods proposed in this study can improve the overall ability of detecting camouflaged targets. The method of pixel-level fusion of visible and IR images produced the best results and the greatest increase. The use of the anchor box optimization algorithm and new loss function also improved the comprehensive detection index of the camouflaged target to a certain extent, which was relatively small. In the process of combining pixel-level fusion with other methods, it also had a positive effect, and its comprehensive detection ability index was improved. The MIF-YOLOv5 algorithm proposed in this study performed admirably in terms of camouflage target detection.

4.4 Comparison experiments

To ensure fairness of the experiment, we trained 50 epochs on the same computer, which was equipped with an AMD R7 6800H CPU and GeForce RTX3060 6 GB GPU, for the testing experiment. We retrained the well-known single-modal target recognition algorithm on the RGB dataset and the IR dataset of the camouflaged target to demonstrate the necessity of fusing the RGB image and IR images of the camouflaged target.

In Table 5, we depict the quantitative outcomes of the performance of the MIF-YOLOv5 algorithm and several advanced single-modal target detection algorithms. MIF-YOLOv5's mAP@0.5:0.95 was much larger than those of other algorithms on the RGB or IR dataset. On the testing set, we confirmed the validity of these strategies. The single-modal target detection algorithm's high missed detection rate and low detection accuracy were evident.

To illustrate the effectiveness of our proposed method, we then compared the MIF-YOLOv5 algorithm with the current two advanced multimodal image fusion target detection methods: SLBAF-Net (Cheng XL et al., 2023) and cross-modality fusion Transformer (CFT) (Fang et al., 2021). These two algorithms were reproduced based on the experimental results in the original papers and were trained on the multispectral dataset used in this study. Table 6 shows the performances of MIF-YOLOv5, SLBAF-Net, and CFT on the verification set. It can be seen that our proposed method still had great advantages in terms of detection accuracy.

Table 5 Performance comparison among various advanced algorithms

Model	Dataset	Number of parameters (×10^6)	Input size	mAP@ 0.5:0.95
Faster R-CNN	RGB	7.5	640×640	0.321
Faster R-CNN	IR	7.5	640×640	0.378
SSD	RGB	6.9	640×640	0.439
SSD	IR	6.9	640×640	0.476
YOLOv3	RGB	6.2	416×416	0.341
YOLOv3	IR	6.2	416×416	0.391
YOLOv4	RGB	6.1	416×416	0.278
YOLOv4	IR	6.1	416×416	0.341
YOLOv5	RGB	6.7	640×640	0.364
YOLOv5	IR	6.7	640×640	0.488
YOLOv8	RGB	10.6	640×640	0.367
YOLOv8	IR	10.6	640×640	0.471
MIF-YOLOv5	RGB+IR	7.0	640×640	0.639

Table 6 Results of multimodal detection algorithms on multispectral datasets

Model	Dataset	Number of parameters (×10^6)	Input size	mAP@ 0.5:0.95
SLBAF-Net	RGB+IR	0.4	640×640	0.207
CFT	RGB+IR	42.8	640×640	0.504
MIF-YOLOv5	RGB+IR	7.0	640×640	0.639

5 Conclusions

In this paper, we presented a novel multispectral camouflaged target dataset and proposed a multimodal image fusion based camouflaged target detection model. Experimental results demonstrated that pixel-level fusion of optical and IR images in the same state space of the camouflaged target can effectively improve the feature information richness of the camouflaged target and fully use the target identification algorithm's potential. Furthermore, the anchor frame optimization and loss function were employed to further optimize the model's convergence process and improve the algorithm's robustness. It is clear that our technology can serve as a model for the quick positioning and high-precision detection of military targets on the battlefield. Note that the multispectral dataset we created is tightly tied to the field test site's vegetation status, topography, illumination conditions, and detector distance. We plan to expand the dataset types in future development.

Contributors

Ruihui PENG and Jie LAI designed the research. Jie LAI devised the experimental method for acquiring camouflage target datasets. Xueting YANG, Yingjuan SONG, and Wei GUO collaborated to complete data collection. Jie LAI and Xueting YANG accomplished experimental verification. Jie LAI and Dianxing SUN drafted the paper. Ruihui PENG, Dianxing SUN, and Shuncheng TAN revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Bhajantri NU, Nagabhushan P, 2006. Camouflage defect identification: a novel approach. Proc 9th Int Conf on Information Technology, p.145-148. https://doi.org/10.1109/ICIT.2006.34

Bochkovskiy A, Wang CY, Liao HY, et al., 2020. YOLOv4: optimal speed and accuracy of object detection. https://arxiv.org/abs/2004.10934

Cheng XL, Geng KK, Wang ZW, et al., 2023. SLBAF-Net: super-lightweight bimodal adaptive fusion network for UAV detection in low recognition environment. *Multim Tools Appl*, 82(30):47773-47792. https://doi.org/10.1007/s11042-023-15333-w

Cheng Y, Hao HZ, Ji Y, et al., 2022. Attention-based neighbor selective aggregation network for camouflaged object detection. Proc Int Joint Conf on Neural Networks, p.1-8. https://doi.org/10.1109/IJCNN55064.2022.9892156

Fan DP, Ji GP, Sun GL, et al., 2020a. Camouflaged object detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2777-2787. https://doi.org/10.1109/CVPR42600.2020.00285

Fan DP, Ji GP, Zhou T, et al., 2020b. PraNet: parallel reverse attention network for polyp segmentation. Proc 23rd Int Conf on Medical Image Computing and Computer-Assisted Intervention, p.263-273. https://doi.org/10.1007/978-3-030-59725-2_26

Fang QY, Han DP, Wang ZK, 2021. Cross-modality fusion Transformer for multispectral object detection. https://arxiv.org/abs/2111.00273

Gevorgyan Z, 2022. SIoU loss: more powerful learning for bounding box regression. https://arxiv.org/abs/2205.12740

Girshick R, 2015. Fast R-CNN. Proc IEEE Int Conf on Computer Vision, p.1440-1448. https://doi.org/10.1109/ICCV.2015.169

Girshick R, Donahue J, Darrell T, et al., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.580-587. https://doi.org/10.1109/CVPR.2014.81

Hu JH, Cui GZ, Qin L, 2015. A new method of multispectral image processing with camouflage effect detection. Proc SPIE 9675, Image Processing and Analysis, Article 967510. https://doi.org/10.1117/12.2199206

Liang XY, Lin HK, Yang H, et al., 2021. Construction of semantic segmentation dataset of camouflage target image. *Lasers Optoelectron Prog*, 58(4):0410015 (in Chinese). https://doi.org/10.3788/LOP202158.0410015

Lin ZY, Goyal P, Girshick R, et al., 2020. Focal loss for dense object detection. *IEEE Trans Patt Anal Mach Intell*, 42(2): 318-327. https://doi.org/10.1109/TPAMI.2018.2858826

Liu CX, 2022. Research on the Fusion Algorithms of Infrared and Visible Image. MS Thesis, Lanzhou Jiaotong University, Lanzhou, China (in Chinese). https://doi.org/10.27205/d.cnki.gltec.2022.001211

Liu W, Anguelov D, Erhan D, et al., 2016. SSD: single shot multibox detector. Proc 14th European Conf on Computer Vision, p.21-37. https://doi.org/10.1007/978-3-319-46448-0_2

Lv YQ, Zhang J, Dai YC, et al., 2021. Simultaneously localize, segment and rank the camouflaged objects. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11591-11601. https://doi.org/10.1109/CVPR46437.2021.01142

Putatunda R, Gangopadhyay A, Erbacher RF, et al., 2022. Camouflaged object detection system at the edge. Proc SPIE 12096, Automatic Target Recognition XXXII, Article 120960I. https://doi.org/10.1117/12.2618869

Qi B, 2022. Research on Fusion of Infrared and Visible Light Image Based on Co-occurrence Analysis Shearlet Transform. MS Thesis, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China (in Chinese). https://doi.org/10.27522/d.cnki.gkcgs.2022.000050

Redmon J, Farhadi A, 2017. YOLO9000: better, faster, stronger. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.7263-7271. https://doi.org/10.1109/CVPR.2017.690

Redmon J, Divvala S, Girshick R, et al., 2016. You only look once: unified, real-time object detection. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.779-788. https://doi.org/10.1109/CVPR.2016.91

Sun XH, Guan Z, Wang X, 2023. Vision Transformer for fusing infrared and visible images in groups. *J Image Graph*, 28(1):166-178 (in Chinese). https://doi.org/10.11834/jig.220515

Tan XY, Hu X, Yang JX, et al., 2022. Camouflaged object detection based on progressive feature enhancement aggregation. *J Comput Appl*, 42(7):2192-2200 (in Chinese). https://doi.org/10.11772/j.issn.1001-9081.2021060900

Wu GJ, Lyu XL, Xing HN, et al., 2015. Application of three-dimensional convex analysis in pattern painting camouflage detection. *J PLA Univ Sci Technol (Nat Sci Ed)*, 16(6): 582-586 (in Chinese). https://doi.org/10.7666/j.issn.1009-3443.20141212001

Yadav D, Arora MK, Tiwari KC, et al., 2018. Detection and identification of camouflaged targets using hyperspectral and LiDAR data. *Def Sci J*, 68(6):540-546. https://doi.org/10.14429/dsj.68.12731

Zhang W, Zhou QK, Li RZ, et al., 2022. Research on camouflaged human target detection based on deep learning. *Comput Intell Neurosci*, 2022:7703444. https://doi.org/10.1155/2022/7703444