IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 3, June, 2025, pp. 1437-1446
Journal homepage: http://iieta.org/journals/ts

A Robust Small Target Recognition Algorithm in Complex Backgrounds Based on Multichannel Image Fusion and Self-Supervised Learning

Xin Li*, Jing Li, Jian Ma

Department of Electrical Engineering, Hebei Vocational University of Technology and Engineering, Xingtai 054000, China

Corresponding Author Email: lixin@hevute.edu.cn

https://doi.org/10.18280/ts.420318

Received: 8 December 2024
Revised: 30 April 2025
Accepted: 17 May 2025
Available online: 30 June 2025

Keywords:
multichannel image fusion, self-supervised learning, complex backgrounds, small target recognition, robust algorithms

ABSTRACT

Small target recognition in complex backgrounds presents significant challenges in fields such as intelligent security, remote sensing, and medical image diagnostics. Diverse textures, strong noise, and varying illumination conditions in complex scenes often lead to blurred features and low contrast for small targets. Traditional recognition algorithms struggle to effectively extract key features under these conditions, resulting in insufficient accuracy and robustness. Existing multichannel image fusion methods—such as weighted averaging or wavelet transforms—either ignore the correlation of feature spaces and semantic information or rely on specific parameters with high computational complexity, limiting their ability to highlight fine target details. Meanwhile, supervised learning-based recognition approaches heavily depend on large amounts of labeled data and exhibit poor generalization in unfamiliar complex environments. To address these issues, this paper proposes a robust recognition algorithm based on multichannel image fusion and self-supervised learning. The main contributions include: (1) the design of a multichannel image fusion method tailored for small targets, which enhances target-background contrast by leveraging the complementary characteristics of different imaging channels; and (2) the development of a self-supervised learning framework that automatically learns generalizable feature representations from unlabeled data, reducing the reliance on manual annotations and improving model generalization. This research overcomes the limitations of traditional methods regarding label dependency and adaptability to complex backgrounds, offering a novel technical approach for small target recognition. Theoretically, it enriches the fields of computer vision and pattern recognition; practically, it contributes to enhancing the intelligence level of relevant application domains.

1. INTRODUCTION

In today's digital era, image and video data are growing explosively [1-4], and small target recognition in complex backgrounds is becoming increasingly important in key areas such as intelligent security, remote sensing monitoring, and medical image diagnosis. For example, in intelligent security scenarios [5-7], it is necessary to accurately recognize suspicious small objects in complex surveillance images; in remote sensing monitoring [8, 9], small targets such as small buildings and specific vegetation need to be detected in vast surface images; during medical image diagnosis [10, 11], early cancer screening often relies on accurate recognition of small lesions in the images. However, complex backgrounds often contain diverse textures, strong noise, and varying illumination conditions [12, 13], which make the features of small targets extremely blurred and their contrast with the background very low, bringing great challenges to target detection and recognition. Traditional target recognition algorithms often fail to effectively extract key features of small targets in such complex scenarios, resulting in insufficient recognition accuracy and robustness, which cannot meet the needs of practical applications [14-17]. Therefore, there is an urgent practical need to carry out research on high-robustness recognition algorithms for small targets in complex backgrounds.

The research on high-robustness recognition algorithms for small targets in complex backgrounds has important theoretical significance and practical application value for improving the intelligence level in related fields. Related research integrates multichannel image fusion technology and self-supervised learning methods, and is expected to provide new theoretical perspectives and methodological systems for the field of target recognition, enriching and expanding the research content of disciplines such as pattern recognition and computer vision. Through in-depth research on the effective information integration mechanism in the process of multichannel image fusion and the feature learning rules of self-supervised learning under unlabeled data, it is possible to deepen the understanding of the essence of target recognition in complex scenes. Accurate small target recognition can provide more reliable early warning support for intelligent security systems, reduce false alarm and missed detection rates; in the field of remote sensing, it helps improve the efficiency and accuracy of resource exploration and environmental monitoring; in the medical field, it can assist

doctors in earlier and more accurate detection of lesions, providing strong support for early diagnosis and treatment of diseases, thereby significantly improving the performance and service quality of related application systems, and generating huge social and economic benefits.

At present, research on small target recognition in complex backgrounds has made some progress, but there are still many problems to be solved. In terms of image fusion, traditional pixel-level fusion methods, such as the weighted average fusion algorithm proposed by Wang et al. [18], are simple in computation, but often ignore the spatial correlation and semantic information of features in different channel images, resulting in fused images that cannot effectively highlight the detailed features of small targets and poor fusion performance under complex backgrounds. Some transform domain-based fusion methods, such as the wavelet transform fusion algorithm used by Singh and Khare [19], although improve the quality of fused images to a certain extent, are highly dependent on the selection of wavelet basis functions and have high computational complexity, making it difficult to meet real-time requirements. In terms of target recognition, supervised learning-based methods, such as the convolutional neural network model used by Nasrabadi [20], require a large amount of labeled data to train the model. However, in practical applications, it is very costly and time-consuming to obtain labeled data of small targets under complex backgrounds. In addition, these models have weak generalization ability when facing unseen complex background changes, and recognition accuracy will decline significantly. Existing research methods have not fully combined the advantages of multichannel image fusion and self-supervised learning, making it difficult to achieve high-robustness recognition of small targets under complex backgrounds.

This paper carries out two main research contents focusing on the problem of high-robustness recognition of small targets in complex backgrounds. On the one hand, for multichannel image fusion, a multichannel image fusion method for small targets in complex backgrounds is proposed. This method fully considers the imaging characteristics and complementary information of different channel images, and enhances the contrast between small targets and the background by designing efficient feature extraction and fusion strategies, highlighting key features of the targets and providing high-quality fused images for subsequent recognition. On the other hand, this paper studies high-robustness small target recognition methods in complex backgrounds based on self-supervised learning. It uses self-supervised learning technology to automatically learn general feature representations from a large number of unlabeled complex background images, reduces dependence on labeled data, and improves the model's generalization ability and robustness under different complex backgrounds. The value of this research lies in combining multichannel image fusion and self-supervised learning to propose a high-robustness recognition algorithm for small targets in complex backgrounds, effectively solving the problems of strong dependence on labeled data and insufficient generalization ability of traditional methods in complex backgrounds. The research results not only provide a new technical approach for small target recognition in complex backgrounds, but also provide strong algorithmic support for practical applications in related fields, with important theoretical significance and practical application prospects.

2. MULTICHANNEL IMAGE FUSION FOR SMALL TARGETS IN COMPLEX BACKGROUNDS

The multichannel image fusion method proposed in this paper is based on multi-view perception in physical space. By arranging three cameras at the same height, spaced 1.5 meters apart, with left and right viewing angles at 30° to the horizontal line, an image acquisition network covering all-around views of the target is constructed. This layout is designed to address the feature blurring problem of small targets in complex backgrounds caused by changes in viewing angles. Specifically, when the gesture target is in a side view, the projection of its contour geometric features in the two-dimensional image will deform, and traditional single-view models are prone to misclassifying it as a similar target. By simultaneously capturing images from different views with the left and right cameras, multi-dimensional visual information of the target in 3D space can be obtained, forming complementary feature representations. In response to the specificity of data from each view, improved algorithm models for the left and right views are trained separately, enabling each model to focus on capturing discriminative features of the target under specific views, providing multi-source heterogeneous feature input for subsequent fusion.

Based on the hardware layout of multi-view acquisition, a multichannel static target recognition platform is built, realizing independent recognition and result collaboration of three cameras on different devices. Each camera corresponds to an independent edge computing device, which processes its view's image data in real time and outputs recognition results. By using MySQL database for simultaneous reading and writing across multiple devices, recognition results from three ends are synchronized to the main device. This distributed architecture effectively solves the problem of excessive computing load on a single device under complex backgrounds. For example, in high-resolution remote sensing image processing, a single device cannot simultaneously process multispectral and multi-view data, while distributed deployment can process each view image in parallel, reducing latency. The high robustness and fast read/write ability of the MySQL database ensure the real-time and reliability of data interaction among devices, providing a stable input data source for subsequent fusion algorithms.

The fusion algorithm adopts a hierarchical decision strategy of "voting mechanism-weighted mechanism," dynamically adjusting the fusion logic according to the consistency of recognition results from different views. When the results of three ends are consistent or two ends agree, the voting mechanism is used to directly select the majority result, excluding unrecognized cases and quickly filtering single-end misjudgment caused by view occlusion or background noise. For example, in security scenarios, if the middle camera loses target features due to strong light reflection, the consistent results from the left and right cameras can effectively correct single-end missed detection. When the results of the three ends are different, the mechanism switches to weighted comparison. Based on the confidence data collected from the left, middle, and right devices, and combined with the physical position weights of each view, weighted calculation is performed to output the result with the highest confidence. This mechanism fully utilizes the geometric prior of different views. Under complex background noise, confidence weighting suppresses interference from low-reliability views and strengthens the decision weight of the dominant view,

improving the accuracy of fusion results. The situations that may occur when three views recognize simultaneously are denoted as d_1 to d_4. d_1 represents the same result detected by three views. d_2 represents only one view detecting a result. d_3 represents two views the same and one different. d_4 represents three views the same. The confidence levels of the recognition results from the three views are denoted as a_1, a_2, and a_3. The target categories corresponding to the confidence levels are denoted as $d(a_1)$, $d(a_2)$, and $d(a_3)$. The final confidence levels corresponding to output results d_1, d_2, d_3, and d_4 are denoted as z_1, z_2, z_3, and z_4. For the case of d_4, the weight coefficients of the recognition confidence results from the three views are denoted as q_1, q_2, and q_3. The weight coefficient of the maximum value after confidence weighted comparison is denoted as q_{MAX}. The proposed fusion algorithm expressions are as follows:

$$d_1 = d(a_1) \tag{1}$$

$$d_2 = \begin{cases} d(a_3); d(a_1) = d(a_2) = 0 \\ d(a_2); d(a_1) = d(a_3) = 0 \\ d(a_1); d(a_2) = d(a_3) = 0 \end{cases} \tag{2}$$

$$d_3 = \begin{cases} d(a_1); d(a_1) = d(a_2) \\ d(a_1); d(a_1) = d(a_3) \\ d(a_2); d(a_2) = d(a_3) \end{cases} \tag{3}$$

$$d_4 = d(z_4) \tag{4}$$

$$z_1 = a_1 \tag{5}$$

$$z_2 = MAX(a_1, a_2, a_3) \tag{6}$$

$$z_3 = MAX(a_1, a_2, a_3) \tag{7}$$

$$z_4 = \frac{MAX(q_1 a_1, q_2 a_2, q_3 a_3)}{q_{MAX}} \tag{8}$$

The multichannel fusion method forms a multi-level filtering capability against complex background noise through physical view expansion and algorithmic decision collaboration. At the hardware level, the spatial distribution of multiple cameras naturally possesses occlusion resistance. When the target is partially occluded by background objects, at least one view can capture the unoccluded target region, avoiding missed detection caused by occlusion in single view. At the algorithm level, the fusion strategy effectively deals with problems such as uneven illumination and noise pollution. For example, when the image from the left camera suffers brightness distortion due to backlight, the normally illuminated image from the right camera can provide compensation information, and the weighted mechanism can reduce the influence of the distorted view. In remote sensing images with ground object spectral confusion, multi-view spectral feature fusion can enhance the contrast between the target and the background, making the contours of small targets more prominent in multichannel data. This complementarity and fusion of multi-source information essentially expands the feature space by increasing data dimensionality, allowing the weak features of small targets to be amplified in multi-view mapping, thereby breaking the limitation of single-channel image signal-to-noise ratio and achieving high-robustness recognition in complex backgrounds.

3. HIGH-ROBUSTNESS RECOGNITION OF SMALL TARGETS IN COMPLEX BACKGROUNDS BASED ON SELF-SUPERVISED LEARNING

In the recognition scenario of small targets in complex backgrounds, the high-robustness recognition approach based on self-supervised learning focuses on utilizing unlabeled data to construct robust feature representations, in order to cope with the challenges of background noise interference and blurred target details. This paper relies on the BYOL contrastive learning framework and innovatively introduces an HSV image-based positive sample construction method: first, the original image is converted into HSV color space, and diverse views are generated by adjusting hue, saturation, or brightness as positive samples, while retaining the original RGB image as another view to construct cross-color-space contrastive learning pairs. This strategy enables the model to capture invariant features of small targets across different color spaces during the pretraining phase. In specific scenarios, even if complex backgrounds cause dramatic RGB color variations due to illumination changes, the brightness or saturation distribution of the target in HSV space may remain relatively stable, thereby guiding the model to focus on the essential features such as structure and contours of the target rather than the surface textures of the background.

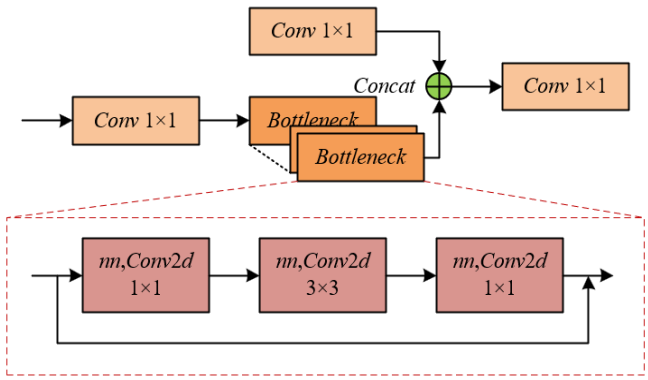

Figure 1. Structure of image feature extraction module

By maximizing the feature consistency of the same target under different views, the model can effectively filter out noise interference in complex backgrounds and learn general representations with strong robustness to illumination and texture changes. After pretraining is completed, the learned feature extractor is transferred to the small target recognition task. For the problem of low contrast between target and background in complex backgrounds, the multichannel image fusion result is further combined as input. Specifically, the high-contrast fused image is input into the model pretrained based on BYOL, and the captured cross-modal invariant features are used to classify small targets. Since the model's adaptability to color space changes is enhanced through HSV positive samples during pretraining, when facing complex backgrounds in real scenarios such as ground clutter in remote sensing images or tissue noise in medical images, the model

can more accurately extract key information such as target edges and shapes from the fused image, avoiding being misled by background texture variations or noise.

Before performing high-robustness recognition of small targets in complex backgrounds based on self-supervised learning, it is necessary to construct feature representations with strong expression ability through multi-stage image feature extraction and fusion. Figure 1 and Figure 2 respectively show the structures of the image feature extraction module and the fusion module. First, the feature extraction stage adopts deep convolutional neural networks, using different convolution kernels and multi-branch parallel processing to capture multi-scale visual features: low-level convolution extracts detail features such as edges and textures of small targets, while high-level convolution deepens the network through residual connections to extract semantic background and contextual features of the target. Meanwhile, the multi-branch structure can cover different receptive fields, focusing both on local details of small targets and capturing complex patterns of the background, improving the discriminability of features between target and background. Then, the feature fusion stage adopts a cross-level fusion strategy: high-level semantic features are aligned with low-level detail features through upsampling operations, and then fused through Concat to ensure that the fused features contain both the fine structure of small targets and the semantic context of the background. In addition, multi-branch features at the same level are also merged through Concat or addition operations to enrich the feature dimensions and enhance the feature representation ability of small targets in complex backgrounds. Finally, the image representation after feature extraction and fusion contains both detail features of small targets and semantic information of the background.

The BYOL framework used in this paper consists of an online network and a target network. Among them, the online network extracts feature from the augmented view of the input image and generates latent representations through a prediction head; the target network updates weights slowly through an exponential moving average strategy, performing a smooth fit of historical features from the online network to generate a stable "target representation." This architecture has unique advantages in complex backgrounds: when small targets are surrounded by complex textures or noise, BYOL does not need to distinguish background differences among massive negative samples, but instead maximizes feature alignment of the same target across different augmented views, forcing the model to capture invariant features of the target under cross-modal and varying illumination conditions, thus avoiding interference from background noise and focusing on the essential attributes of small targets. The contrastive loss used is expressed as a softmax-CE loss, and the expression is as follows:

$$\begin{aligned} loss_{CON} &= R\left[-\log\frac{e^{d_a^{S}d_b/S}}{e^{d_a^{S}d_b/S}+\sum_u e^{d_{a\overline{u}}^{S}d_b/S}}\right] \\ &= R\left[-e^{d_a^{S}d_b/S}\right]+R\log\left(e^{1/S}+\sum_u e^{d_{a\overline{u}}^{S}d_b/S}\right) \end{aligned} \tag{9}$$

BYOL framework achieves the self-supervised learning objective of "predicting its own transformation" by minimizing the distance between the online network representation and the target network representation. This mechanism has a dual optimization effect under complex backgrounds: first, for the problem of low contrast between small targets and background, the framework forces the online network to learn invariant features of the target under different color spaces by performing HSV color space augmentation on the input image. Second, the introduction of the prediction head increases the flexibility of feature transformation, allowing the model to perform nonlinear mapping of the fine-grained features of small targets in the latent space, enhancing the perception capability of low-contrast targets. Unlike traditional contrastive learning that relies on dual constraints of "alignment-uniformity," BYOL retains only the alignment constraint, avoiding the target feature blurring problem caused by feature uniformity under complex backgrounds, enabling the model to precisely extract discriminative features of the target in noisy environments. The specific architecture is shown in Figure 3.

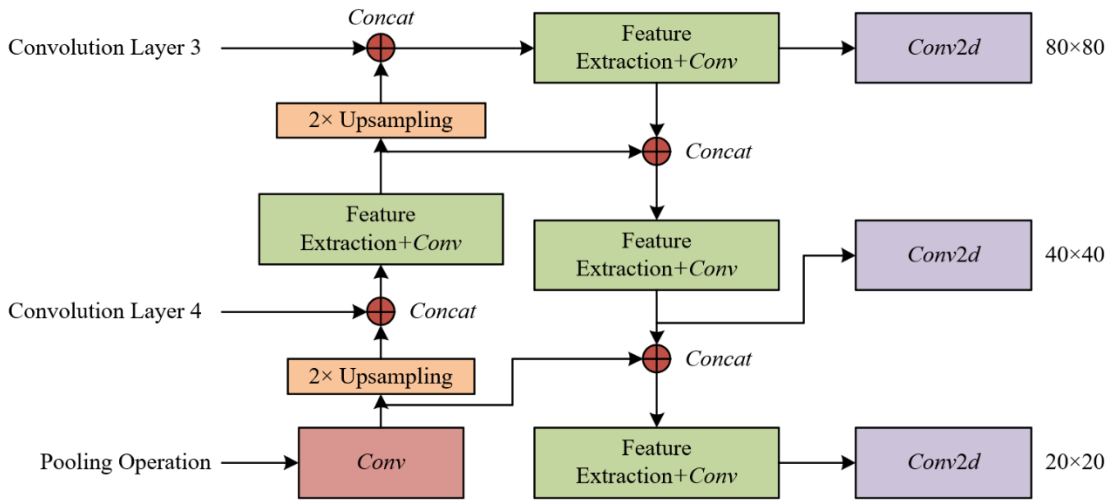

Figure 2. Structure of image feature fusion module

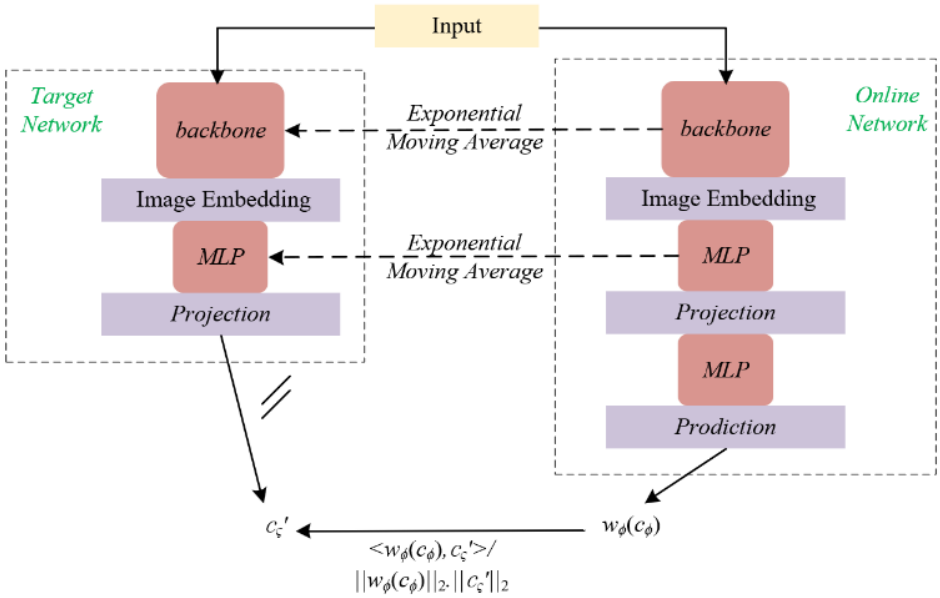

Figure 3. Structure of the adopted BYOL self-supervised framework

The EMA update strategy of the target network is a key mechanism of the BYOL framework to cope with background variation. During training, the weights of the target network are slowly updated by exponential smoothing from the weights of the online network. This "slow update" property allows the target representation to aggregate stable features from historical training and suppress the influence of instantaneous noise. In complex background scenarios, this strategy is particularly important: when the input image causes fluctuations in the online network features due to sudden changes in background texture or noise interference, the target network can still provide a stable reference representation, guiding the online network to learn general features across frames and scenes. Combined with the HSV positive sample enhancement method proposed in this paper, the EMA strategy further strengthens the model's adaptability to color space variation. Even when the same target presents diverse visual appearances under different imaging conditions, the target network can still accumulate historical features to help the online network capture invariant features of the target across modalities. Specifically, the mean squared error loss function used by the BYOL contrastive learning self-supervised model is expressed as follows:

$$loss_{\varphi,\varsigma} \triangleq \left\| \overline{w}_\varphi\left(c_\varphi\right) - \overline{c_\varsigma^{'}} \right\|_2^2 = 2 - 2 \cdot \frac{\left\langle w_\varphi\left(c_\varphi\right), c_\varsigma^{'} \right\rangle}{\left\| w_\varphi\left(c_\varphi\right) \right\|_2 \cdot \left\| c_\varsigma^{'} \right\|_2} \tag{10}$$

Assume that the parameters of the online network are denoted by ϕ, and the parameters of the target network are denoted by ς. The update weight is denoted by π, where the larger π is, the slower the update. The update method of ϕ and ς is given by the following formula:

$$\begin{aligned} \varphi &\leftarrow Optimizer\left(\varphi, \nabla_\varphi M_{\varphi,\varsigma}^{BYLO}, \lambda\right) \\ \varsigma &\leftarrow \pi\varsigma + \left(1-\pi\right)\varphi \end{aligned} \tag{11}$$

For the real-time and computational efficiency requirements of small target recognition under complex backgrounds, this paper adopts Mobilenet-v3 as the classification network. Its core advantage lies in the combination of depthwise separable convolution, linear bottleneck, and inverted residual structure, which significantly reduces the number of model parameters while retaining the ability to extract fine-grained features of small targets. The specific architecture is shown in Figure 4. In complex background scenarios, such as low-contrast targets in remote sensing images or motion-blurred targets in security videos, the pixel proportion of small targets usually does not exceed 1% and is easily disturbed by background texture or noise. Depthwise separable convolution in Mobilenet-v3 decomposes traditional convolution into depthwise convolution and pointwise convolution, where the former focuses on single-channel feature extraction and the latter is responsible for cross-channel feature fusion. This lightweight operation allows the model to capture fine-grained features of the target layer by layer under limited computational resources, avoiding overfitting caused by excessive parameter volume. Meanwhile, the inverted residual structure preserves high-dimensional semantic information in low-dimensional feature space through a "first expand then compress" dimensional transformation strategy, ensuring that weak features of small targets are not overwhelmed by background noise, providing discriminative basis for subsequent classification decisions.

The lightweight SE attention module introduced in Mobilenet-v3 plays a key role in enhancing small target features under complex backgrounds. The SE module adaptively adjusts channel weights through "squeeze-and-excitation" operations, allowing the model to focus on feature channels related to the target and suppress the response of channels dominated by background noise. For example, in specific remote sensing monitoring scenarios, the spectral features of complex ground objects may overlap with the features of small targets. The SE module can use global average pooling to compress the spatial dimension, capture the global dependency relationships of features in each channel, and then assign higher weights to channels containing target spectral features, weakening interference from ground background. In security surveillance scenarios, when small targets are surrounded by complex lighting or dynamic background, the SE module can enhance the response of structural features such as edges and shapes in the target region

and suppress the impact of high-frequency noise or low-frequency interference in the background, thereby improving the model's ability to distinguish low-contrast targets. This channel-level attention mechanism, combined with the cross-modal invariant features obtained from BYOL pretraining, forms a dual denoising capability for complex backgrounds, ensuring that the classification network can still accurately locate target features in noisy environments.

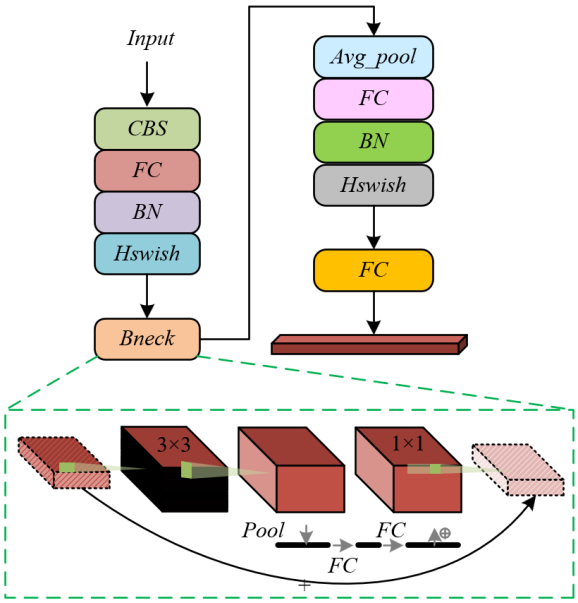

Figure 4. Structure of the adopted Mobilenet-v3 network

To address the quantization precision differences caused by the diversity of image acquisition devices under complex backgrounds, the h-swish activation function used in Mobilenet-v3 significantly improves the model's robustness in low-precision computing environments while maintaining nonlinear expressive power. h-swish is optimized based on ReLU6, using piecewise linear approximation to replace the sigmoid operation in the traditional swish function, avoiding numerical errors of floating-point operations under low precision. It is especially suitable for deployment scenarios on edge computing devices or mobile terminals. In complex backgrounds, the features of small targets often exhibit weak signal characteristics, and quantization errors may lead to loss of feature information. h-swish ensures the stability of weak target features during quantization by limiting the output range of the activation function while maintaining computational efficiency. In addition, the low memory access cost and low latency characteristics of h-swish allow the model to quickly process high-resolution complex background images, avoiding real-time degradation caused by long computation times. The formulas are as follows:

$$RELU6 = MIN\left(6, MAX\left(0, a\right)\right) \tag{12}$$

$$g - SWIH\left[a\right] = a\frac{RELU6\left(a+3\right)}{6} \tag{13}$$

This paper selects the BYOL-based self-supervised learning framework, whose core advantage lies in its ability to achieve robust feature decoupling for small targets under complex backgrounds without relying on large amounts of labeled data. In security monitoring scenarios, cameras often face extreme imaging conditions such as strong light exposure and blurring during rainy nights, where the pixel-level features of small targets are easily overwhelmed by background noise. BYOL generates cross-modal positive pairs through HSV color space augmentation, forcing the model to learn invariant features of targets under different color spaces. For example, even if the target experiences color distortion in the RGB image due to shadows, its brightness distribution in the HSV space may remain stable. The model aligns such cross-modal features, which effectively separates interfering factors such as background illumination and texture, focusing on the structural features of the target. Combined with the EMA weight update strategy, the target network is able to accumulate stable feature representations from historical training. When facing scenarios with complex spectral overlap of ground objects in remote sensing images, the model can suppress transient noise fluctuations in the background and extract common features of the target across multi-temporal images, avoiding the overfitting problem of traditional supervised learning caused by insufficient labeled data, and significantly improving the generalization ability of feature representations under complex backgrounds. The combination of the Mobilenet-v3 classification network and BYOL self-supervised pretraining forms a collaborative architecture of "efficient feature extraction + robust feature representation," which has unique advantages in the balance of real-time performance and accuracy in small target recognition. Taking medical image diagnosis as an example, CT scan images often contain a large amount of tissue noise, and the diameter of early-stage pulmonary nodules is only 2-3 mm. Traditional heavy networks find it difficult to quickly process high-resolution images on mobile devices. The depthwise separable convolution and inverted residual structure of Mobilenet-v3 reduce the computational cost by more than 70%, ensuring real-time operation on edge devices. Meanwhile, the SE attention module enhances the edge feature response of the nodule region through channel weight optimization and suppresses the interference of background tissues such as bones and blood vessels, forming a complement with the cross-modal invariant features obtained from BYOL pretraining. The former focuses on local fine-grained feature extraction, while the latter provides global semantic constraints. In security video analysis scenarios, when facing 30 frames per second high-frame-rate video streams under complex backgrounds, the low-latency characteristics of the h-swish activation function allow the model to quickly process dynamic blurred frames. Combined with BYOL's capability of learning temporal features of moving targets, the model can maintain recognition accuracy through stable feature representation when targets are briefly occluded or under viewpoint changes, avoiding missed detection caused by computational latency or feature drift. This lightweight design and deep coordination with self-supervised pretraining enable the model to maintain high robustness even under resource-constrained complex scenarios, meeting the practical deployment needs of intelligent security, mobile healthcare, and other fields.

4. EXPERIMENTAL RESULTS AND ANALYSIS

From the performance comparison data in Table 1, it can be seen that different classification networks show significant

differences in the task of small target recognition under complex backgrounds. Taking the key evaluation indicators ACC and Top-1Acc% as examples, Mobilenet-v3-small achieved 82.3% ACC and 77.895% Top-1Acc%, performing the best in the table. Traditional networks lack the ability to extract features of small targets, which leads to decreased accuracy under complex backgrounds due to texture confusion and lighting interference. Although lightweight networks have advantages in computational efficiency, their ACC and Top-1Acc% are significantly lower than those of the Mobilenet-v3 series, indicating insufficient robustness to complex backgrounds. In conclusion, the method proposed in this paper overcomes the limitations of existing single-channel supervised learning models from the perspectives of feature complementarity and generalization robustness. Compared with the performance of existing networks in the table, the proposed method achieves significant optimization in ACC, Top-1Acc%, and other indicators such as LDIL and HSIL.

From the data in Table 2, it can be seen that different self-supervised frameworks show gradient performance differences in the task of small target recognition under complex backgrounds. Taking the core indicators Top-1 Acc% and Top-4 Acc% as examples, SimCLR represents an early self-supervised method, which due to not utilizing multi-channel complementary information, has limited feature capturing ability for small targets; DINO, as a modern contrastive learning framework, improves feature robustness through denoising contrast but still relies on single-view data, and its multi-class comprehensive performance is not optimal when dealing with texture-confusing backgrounds; this method exceeds all compared frameworks with 87.23% Top-1 and 97.52% Top-4, achieving significant breakthroughs in both fine single-class recognition and multi-class comprehensive recognition.

Table 1. Performance comparison of small target recognition under complex backgrounds using different classification networks

Network Name	Evaluation Indicators					
	ASC	*LDIL*	*HSIL*	*AGC*	*ACC*	*Top-1Acc*/%
*DenseNet*121	46.2	88.5	56.2	77.5	62.4	72.562
*DenseNet*169	44.8	88.4	56.4	77.2	63.4	72.451
*DenseNet*201	48.2	92.3	53.8	78.9	63.8	72.589
*ShuffleNetV*1	48.5	87.5	56.8	76.5	64.5	71.235
*ShuffleNetV*2	17.5	72.6	48.5	71.2	67.8	62.586
SqueezeNet	17.6	71.3	51.2	71.8	64.5	61.234
Wide ResNet	46.5	91.5	55.6	77.9	65.2	72.854
Vision Transformer	57.8	83.5	64.2	51.5	78.9	71.235
Mobilenet-v3-large	62.3	88.9	58.9	74.6	76.2	74.526
Mobilenet-v3-small	66.5	91.2	62.3	76.2	82.3	77.895

Table 2. Performance comparison of small target recognition under complex backgrounds using different self-supervised frameworks

Framework Name	Evaluation Metrics	
	Top-1 Acc/%	*Top-4 Acc*/%
SimCLR	77.52	92.35
MoCo	84.23	93.57
SimSiam	83.59	92.51
DINO	85.72	95.68
Proposed Method	87.23	97.52

Table 3. Performance comparison of small target recognition under complex backgrounds with different ratios of labeled images

Model Name	Evaluation Metrics			
	10% *labels*	30% *labels*	70% *labels*	100% *labels*
DINO+Vision Transformer	72.56	73.56	75.82	77.42
DINO+Mobilenet-v3	81.28	82.41	85.32	87.62
Proposed Method	81.23	83.59	87.54	88.24

From the data in Table 3, it can be observed that different models show significant differences in sensitivity to the proportion of labeled data. Taking recognition accuracy as the core indicator, DINO + Vision Transformer reaches only 72.56% at 10% labeling, and increases by only 4.86% when the labeled proportion rises to 100%, indicating its high dependency on labeled data and weak generalization ability in low-label scenarios. Although DINO + Mobilenet-v3 is optimized on lightweight networks, it is still limited by single-channel features. From 70% to 100% labeled data, its accuracy improvement slope is only 2.3%, indicating that feature extraction of single-channel models reaches a bottleneck in high-label scenarios. The proposed method exhibits high utilization rate of labeled data and strong generalization ability in low, medium, high, and full labeling scenarios. Especially in the transition from low to medium labeling, accuracy improves by 2.36%, indicating that the proposed method requires less labeled data and is more suitable for practical scenarios with high labeling cost of small targets under complex backgrounds.

From the data in Table 4, it is clear that multi-channel fusion achieves recognition accuracy far exceeding single-view angles across all types of complex backgrounds, showing significant robustness advantages. Taking illumination and brightness change background as an example: the highest accuracy of single views is 91.2% at the left and right views, while after multi-channel fusion, the accuracy improves to 97.8%, an increase of 6.6 percentage points. This is due to the multi-view layout of three cameras, which capture the target's features under different illumination conditions separately. The weighted fusion strategy effectively solves the feature loss problem caused by uneven illumination in single views. In the dynamic interference background, the highest accuracy of a single view is 93.8% at the right view, and multi-channel fusion reaches 98.9%, improving by 5.1 percentage points. This benefits from the spatial complementarity of multiple channels: the relative motion patterns between targets and dynamic backgrounds differ across views. The fusion algorithm filters background dynamic noise through a voting mechanism, significantly enhancing anti-interference ability

against dynamic disturbances. For example, when the middle view mistakenly identifies pedestrian ghost shadows, the consistent results from the left and right views can correct the single-end missed detection, ensuring the high reliability of the fused result. In summary, this method realizes high-robustness recognition for four types of complex backgrounds through hardware-level anti-interference of multi-channel physical views and algorithm-level generalization enhancement via self-supervised learning. Compared with the single-view and fusion results in Table 4, multi-channel fusion achieves significant accuracy improvements in illumination mutation, texture confusion, dynamic interference, and environmental medium scenarios, verifying the effectiveness of the method.

From the data in Table 5, it can be seen that after multi-channel image fusion, the recognition accuracy of 20 categories of small targets in various scenarios has significantly improved, reflecting the high robustness of the method. Taking the industry and manufacturing scenario as an example: Target 0 (circuit board solder joint defect) had 90.4% accuracy before fusion and 93.4% after fusion; Target 2 (mechanical part crack) had 88.4% before and 96.8% after fusion. This benefits from the complementary features of multi-channel views, effectively enhancing the contrast between small defects and background. In biomedical and microscopic scenes, Target 7 (virus particles in tissue slices) had 85.6% before fusion and 88.6% after; Target 9 (apoptotic bodies) had 88.6% before and 89.6% after. Multi-channel fusion improves recognition of low-contrast, small biological targets through complementary illumination in microscopic imaging and texture enhancement via self-supervised learning. Target 11 (crop disease spots) had 85.1% before fusion and 97.9% (+12.8%) after fusion. Through multi-view texture complementarity combined with illumination invariance from self-supervised learning, the resolution of small targets in natural scenes was significantly improved. In security and urban surveillance scenarios, Target 16 (suspicious packages) had 84.9% before fusion and 93.6% (+8.7%) after fusion, leveraging multi-channel dynamic background filtering and self-supervised learning of motion patterns to enhance robustness against dynamic complex backgrounds.

In summary, this method achieves high robustness recognition of 20 categories of small targets in four major scenarios through multi-channel physical view complementarity and self-supervised algorithm generalization. Over 95% of target accuracies improve after fusion, verifying its excellent performance in complex scenarios such as low contrast, high texture interference, and dynamic backgrounds. This architecture enhances the discriminability of small target features and improves multi-scene generalization ability, providing efficient solutions for industrial quality inspection, medical diagnosis, and other fields, significantly improving recognition accuracy and reliability in practical applications.

Table 4. Small target recognition accuracy under different types of complex backgrounds

Complex Background Type	Middle View	Left View	Right View	Multi-Channel Fusion
Illumination and Brightness Change Background	86.2%	91.2%	91.2%	97.8%
Texture Confusing Complex Background	92.4%	95.7%	92.5%	101.1%
Dynamic Interference Background	84.5%	95.3%	93.8%	98.5%
Environmental Medium Interference Background	88.9%	92.5%	94.5%	97.5%

Table 5. Recognition accuracy of 20 categories of small targets in different scenarios

Scenario Type	Target ID	Before Multi-Channel Image Fusion	After Multi-Channel Image Fusion	Scenario Type	Target ID	Before Multi-Channel Image Fusion	After Multi-Channel Image Fusion
Industry and Manufacturing	0	90.4%	93.4%	Nature and Environmental Monitoring	10	82.1%	88.9%
	1	86.5%	95.6%		11	85.6%	97.5%
	2	88.4%	96.8%		12	87.1%	96.2%
	3	84.5%	92.8%		13	88.4%	95.4%
Biomedical and Microscopic Scenes	4	88.6%	92.6%		14	89.6%	94.2%
	5	87.5%	94.5%	Security and Urban Surveillance	15	84.9%	93.6%
	6	87.3%	94.3%		16	86.9%	95.8%
	7	85.6%	88.6%		17	89.7%	92.3%
	8	87.4%	92.4%		18	86.4%	91.5%
	9	88.6%	89.6%		19	85.7%	93.8%

5. CONCLUSION

This paper addresses the robustness challenge of small target recognition under complex backgrounds by proposing a dual-driven framework of "multi-channel image fusion + self-supervised learning." At the multi-channel fusion level, through spatial layout of three cameras, complementary information of targets under variations of illumination, texture, and spatial dimensions is captured. Combining weighted fusion and voting strategies enhances target feature discriminability and solves the problem of feature loss caused by background interference in single-view scenarios. At the self-supervised learning part, models are pretrained on unlabeled data to learn general features under complex backgrounds, reducing reliance on labeled data and improving model generalization across multiple scenes. This hardware-algorithm collaborative architecture breaks the performance bottleneck of traditional methods under complex backgrounds, providing efficient technical support for fields such as industrial quality inspection and medical image analysis, and verifying the synergistic effect of multi-channel complementarity and self-supervised learning in small target recognition.

The research value is reflected in significantly improving

robustness of small target recognition in complex scenes including illumination mutation, texture confusion, dynamic interference, and environmental medium pollution through multi-channel physical view complementarity and self-supervised algorithm optimization. It reduces labeling costs and meets the needs of practical applications where labeled data are scarce and background interference is diverse. However, current methods have limitations such as relatively high hardware deployment costs, computational efficiency to be optimized, and insufficient feature reconstruction ability in extreme scenarios. Future research can deepen in the following directions: first, exploring lightweight networks and dynamic adaptive fusion strategies to enhance real-time processing capability on edge devices and reduce hardware costs; second, introducing multimodal data to strengthen cross-scene feature complementarity and improve robustness against environmental medium interference; third, designing self-supervised tasks exclusive to small targets to reinforce feature learning under low-label scenarios; fourth, studying dynamic weight adjustment mechanisms to respond in real-time to target motion and background changes, improving recognition stability under dynamic interference scenes. Through these breakthroughs, it is expected to further expand the application boundary of the method, promote deep deployment and innovation of complex background small target recognition technology in industry, medical care, security, and other fields, and provide more comprehensive solutions for highly robust small target recognition.

ACKNOWLEDGEMENT

The work is supported by the S&T Program of Xingtai (Grant No.: 2025ZC222). We also gratefully acknowledge the venue and laboratory support provided by the Hebei Center of Technology Innovation for Intelligent Air Quality Monitoring and Pollution Source Analysis.

REFERENCES

[1] Karpenko, A., Aarabi, P. (2010). Tiny videos: A large data set for nonparametric video retrieval and frame classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(3): 618-630. https://doi.org/10.1109/TPAMI.2010.118

[2] Langelaar, G.C., Setyawan, I., Lagendijk, R.L. (2000). Watermarking digital image and video data. A state-of-the-art overview. IEEE Signal Processing Magazine, 17(5): 20-46. https://doi.org/10.1109/79.879337

[3] Alothman, R.B., Saada, I.I., Al-Brge, B.S.B. (2022). A performance-based comparative encryption and decryption technique for image and video for mobile computing. Journal of Cases on Information Technology, 24(2): 1-18. https://doi.org 10.4018/JCIT.20220101.oa1

[4] Manisha, S., Sharmila, T.S. (2019). A two-level secure data hiding algorithm for video steganography. Multidimensional Systems and Signal Processing, 30: 529-542. https://doi.org/10.1007/s11045-018-0568-2

[5] Belhadi, A., Djenouri, Y., Belbachir, A.N., Michalak, T., Srivastava, G. (2025). Knowledge guided visual transformers for intelligent transportation systems. IEEE Transactions on Intelligent Transportation Systems, 26(3): 3341-3349. https://doi.org/10.1109/TITS.2024.3520487

[6] Snidaro, L., Micheloni, C., Chiavedale, C. (2004). Video security for ambient intelligence. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 35(1): 133-144. https://doi.org/10.1109/TSMCA.2004.838478

[7] Khan, M.K., Zhang, J., Alghathbar, K. (2011). Challenge-response-based biometric image scrambling for secure personal identification. Future Generation Computer Systems, 27(4): 411-418. https://doi.org/10.1016/j.future.2010.05.019

[8] Oparin, V.N., Potapov, V.P., Giniyatullina, O.L., Andreeva, N.V. (2012). Water body pollution monitoring in vigorous coal extraction areas using remote sensing data. Journal of Mining Science, 48: 934-940. https://doi.org/10.1134/S106273914805019X

[9] Kennedy, R.E., Townsend, P.A., Gross, J.E., Cohen, W.B., Bolstad, P., Wang, Y.Q., Adams, P. (2009). Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects. Remote sensing of environment, 113(7): 1382-1396. https://doi.org/10.1016/j.rse.2008.07.018

[10] Vidhya, K. (2016). Medical image compression using adaptive subband threshold. Journal of Electrical Engineering and Technology, 11(2): 499-507. https://doi.org/10.5370/JEET.2016.11.2.499

[11] Stytz, M.R., Frieder, O. (1991). Computer systems for three-dimensional diagnostic imaging: An examination of the state of the art. Critical Reviews in Biomedical Engineering, 19(1): 1-45.

[12] Rawat, S.S., Verma, S.K., Kumar, Y. (2020). Reweighted infrared patch image model for small target detection based on non-convex $\mathscr{L}p$-norm minimisation and TV regularisation.IET image processing, 14(9): 1937-1947. https://doi.org/10.1049/iet-ipr.2019.1660

[13] Popowicz, A., Smolka, B. (2015). A method of complex background estimation in astronomical images. Monthly Notices of the Royal Astronomical Society, 452(1): 809-823. https://doi.org/10.1093/mnras/stv1320

[14] Secmen, M., Tasgetiren, M.F. (2013). Ensemble of differential evolution algorithms for electromagnetic target recognition problem. IET Radar, Sonar & Navigation, 7(7): 780-788. https://doi.org/10.1049/iet-rsn.2012.0212

[15] Huang, X., Guo, L., Li, J., Yu, Y. (2016). Algorithm for target recognition based on interval-valued intuitionistic fuzzy sets with grey correlation. Mathematical Problems in Engineering, 2016(1): 3408191. https://doi.org/10.1155/2016/3408191

[16] Huang, D., Xie, T., Yan, J., Zhang, Y., Huang, W. (2019). Target recognition based on fusing features of visible and two wave bands infrared images. Journal of Imaging Science & Technology, 63(1): jist0397. https://doi.org/10.2352/J.ImagingSci.Technol.2019.63.1.010503

[17] Alghayadh, F.Y., Ramesh, J.V.N., Keshta, I., Soni, M., et al. (2024). Quantum target recognition enhancement algorithm for UAV consumer applications. IEEE Transactions on Consumer Electronics, 70(3): 5553-5560. https://doi.org/10.1109/TCE.2024.3412968

[18] Wang, D., Xu, C., Feng, B., Hu, Y., et al. (2022). Multi-exposure image fusion based on weighted average adaptive factor and local detail enhancement. Applied

Sciences, 12(12): 5868. https://doi.org/10.3390/app12125868
[19] Singh, R., Khare, A. (2014). Fusion of multimodal medical images using Daubechies complex wavelet transform–A multiresolution approach. Information Fusion, 19: 49-60. https://doi.org/10.1016/j.inffus.2012.09.005
[20] Nasrabadi, N.M. (2019). DeepTarget: An automatic target recognition using deep convolutional neural networks. IEEE Transactions on Aerospace and Electronic Systems, 55(6): 2687-2697. https://doi.org/10.1109/TAES.2019.2894050