Neurocomputing 549 (2023) 126466

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Extraordinary MHNet: Military high-level camouflage object detection network and dataset

Maozhen Liu, Xiaoguang Di *

Control and Simulation Center, Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Article history:
Received 1 March 2023
Revised 6 June 2023
Accepted 11 June 2023
Available online 14 June 2023
Communicated by Zidong Wang

Keywords:
High-level military camouflage
Object detection
Concealed objects
Datasets

ABSTRACT

We present the first systematic work on Military High-level Camouflage object Detection (MHCD), aiming to identify objects visibly embedded in chaotic backgrounds. The high intrinsic similarities (e.g., texture, intensity, color, etc.) between the attention object and its background give the task far more challenging than general object detection. In this paper, we construct a benchmark MHCD2022 dataset, which consists of 3000 images with dense annotations covering 5 categories from multiple real-world scenes. Remarkably, based on the observation that biological vision usually first obtains perception from global search and strives to recover the complete object, we propose a novel Military High-level detection Network, called MHNet, which is characterized by four ingenious modules: Subject Perception Gathering (SPG), Part-object Relationships Mining (PRM), Concept Recovery/Feature Clue Supplement (CR/FCS) and Springboard Selection (SS). Firstly, a SPG is designed for global foreground rough perception by the exploitation of depth information. Second, a PRM is particularly used to mine part-object potential relations in diverse environments. After that, we propose CR/FCS and SS to enhance the destroyed instance-level representation and suppress the domain imbalance problem, respectively. Extensive experimental results show that previous methods suffered from poor performance, MHNet significantly outperforms camouflage baselines and competing methods on the MHCD2022 for the high-level camouflaged object. Finally, we also present and highlight the practical application value and several future directions of the research.

1. Introduction

Can you quickly find and locate the high-level camouflaged military objects from each image in Fig. 1? We define objects that are difficult for the human eye to immediately discern as high-level camouflage and those that can be quickly distinguished as the lower-level. Given an arbitrary image, the goal of general object detection is to determine predefined classes of presences and locate them in this image. In contrast, identifying concealed/ camouflaged objects in a biological vision scene becomes more intractable. Those objects are masters who are concealed in the surrounding environment. They have other appearances that imitate the same body colors, texture, pattern, and other backgrounds, such as dazzle camouflage [1], camouflage [2], disruptive coloration [3], and distractive markings [4]. This misleading information hinders the most discriminating feature of network learning camouflage. We believe that more visual perception knowledge/ prior about camouflage objects is beneficial.

Recently, camouflage patterns have been widely used in military equipment, and the technique makes it harder to efficiently distinguish an object's appearance from a concealed background, as the foreground object has a very similar appearance to its surroundings. Visual cues used for identification such as texture, edges, color, contrast, and scale are vulnerable to attacks by basic camouflage strategies e.g. disruptive coloration. Closely related work brings convenience to the decipherment of high-level camouflage techniques for warfighter or UAV reconnaissance. To the best of our knowledge, military camouflage object detection has not been studied in the literature before, especially for high-level military camouflage patterns. One of the main reasons is that no accessible object detection benchmark has so far favored the military domain for high-level camouflage. As a result, previous detection methods with military objectives have been mainly grounded in conventional operations such as multi-sensor information processing [5] and migration learning [6], and readers have rarely seen works that directly use deep learning. From the perspective of visual attention mechanisms, as a close task, the relationship

* Corresponding author.
E-mail address: dixiaoguang@hit.edu.cn (X. Di).

https://doi.org/10.1016/j.neucom.2023.126466
0925-2312/

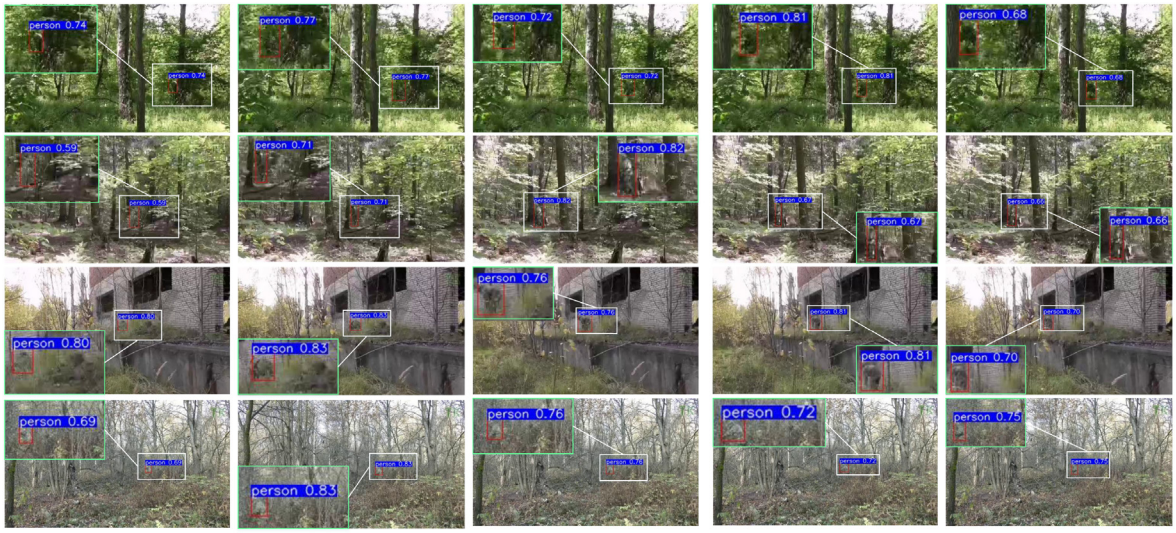

Fig. 1. Some samples from the constructed high-level military camouflage dataset.

between saliency object detection (SOD) and camouflage recognition is orthogonal and adversarial, with SOD working to distinguish the most attention-grabbing regions rather than MHCD ought to reduce the initial saliency level of foreground objects, i.e. the game theory of conspicuous and unseen objectives. Here, it is clearly beyond the scope of this paper to provide an adequate review of the SOD, and we encourage the reader to refer to recent research work [7,8] for the desired details. The suitability of the above approach is tenuous in the concealed context, we aim to explore prior contributions for high-level camouflaged object detection. Moreover, other hardworking researchers have proposed some common methods for infrared and visible image fusion [50–53]. In particular, Tang L et al. [50] proposed a semantic-aware real-time image fusion network (SeAFusion) by bridging the gap between image fusion and high-level vision tasks. SuperFusion [52] incorporates image registration, image fusion, and semantic requirements of high-level vision tasks into a single framework, and achieves satisfactory experimental results. Visible and infrared fusion approaches will be the focus of our future research efforts due to the reconciliation of the mutual advantages and development potential of both. As one of the cores of vision fundamentals, we only discuss visible-image-based approaches for deciphering tasks at this stage in this paper and will not repeat them here.

Based on the observation of biological vision, predators initially perceive “abruptness” (i.e., inconsistency) from a global scene with patterns almost identical to the object and pop up possible rough regions when searching for camouflaged prey; Following this, the predator carefully analyzes those regions in which possible objects and potential parts of the body are based on the a priori knowledge learned from previous stages; further treatments are similar to the instances-level active recovery or optimization of the camouflaged objects. As such, those camouflaged objects can be correctly discriminated against. However, the perfect treatment of the MHCD problem faces several pitfalls as follows: (1) the actual military environment is more complex and the object scale is more variable. (2) Blurred background. (3) High intrinsic similarities of textures between the body and the background. (4) Intra-class differences, same-class objects such as persons with specific patterns (e.g. weed cloak or snowsuit) may be very different. (5) Inter-class similarity, objects belonging to different classes may be very similar, e.g. tanks and military vehicles. In general, military high-level camouflage patterns are usually focused on more cluttered environments such as battlefields, jungles, towns, or mountains. The lighting and obscuration situations are extremely complex and harsh. Therefore, it is difficult to be identified by general detection methods.

To contribute to the development of this research area, we propose a novel MHNet framework and construct a dataset in an attempt to deal with the problem of discrimination in the military context, under high-level camouflage scenes. MHNet consists of four intuitive modules: Subject Perception Gathering (SPG), Part-object Relationships Mining (PRM), Concept Recovery/Feature Clue Supplement (CR/FCS), and Springboard Selection (SS). To capture the global ”abrupt” information representation, SPG is specially designed to generate a depth feature architecture similar to the geoscientific ”contour topography” by using shallow features containing rich boundary information and detail information. Inspired by promising results [9], we took the initiative to introduce such a process of part and object potential property mining into MHCD, considering the aspect that most previous methods ignore that an object normally consists of several related parts combined, and related parts can form the whole object. Next, we propose a CR/FCS component with reverse attention for rescuing or recovering cues interrupted by obstacles such as occlusion or local feature presentation, which is similar to the ”brain patching” process in human behavior. In addition, to cope with the problem of multi-source data, where inter-domain inconsistency of samples is analyzed, we propose a simple and efficient SS module to stabilize the model in the case of hard and easy samples with an unfavorable skew towards the latter, which tends to cause the network to fail in learning hard samples. The main contributions of this paper are listed as follows:

(1) Based on careful bio-visual observation, we propose a novel Military High-level detection Network, called MHNet, for military high-level camouflage object detection. It will facilitate the development of individual combat systems on the real battlefield to achieve rapid object detection and preemptive strike, which is a favorable basis for motivating new ideas on this task.
(2) We design a global ”abrupt” perceptual strategy SPG module to obtain subject-aware a priori by constructing a depth representation for objects, which can discover the agnostic coarse location of the object and ameliorate the weak boundary problem.
(3) We propose a PRM module that integrates part-object relations based on the knowledge of SPG and deep feature encoding learning to enhance the capture of the whole

object by the model with the supplementary part. To our knowledge, this is the first preliminary attempt to incorporate part-object relationship mining into military high-level camouflage object detection.

(4) To mimic human brain activity, we further propose a concept recovery strategy, termed CR/FCS, to repair the destroyed features by fusing reverse attention search with boundary features. In addition, the SS method is cleverly designed to alleviate the domain inconsistency problem brought by training on multiple sources of data and successfully forces the model to take into account the stable learning of "hard, easy" samples.

(5) We created a solid benchmark, the Military High-level Camouflage Dataset MHCD2022, to contribute significantly to the future research process. Through the experiments in the MHCD2022, the proposed MHNet method greatly outperforms several advanced baselines and has significant advantages and friendliness in the identification of high-level camouflaged objects compared with the state-of-the-art methods.

2. Related work

Types of Camouflage. Camouflaged samples can be roughly divided into two types, i.e. natural camouflage and artificial camouflage. Natural camouflage is used by animals as survival skills, such as seahorses, starfish, and chameleons, to confuse predators. Furthermore, artificial camouflage is often used in commercial designs/games to hide information. It appears in life products during the manufacturing process, or in military equipment (the kind we care about).

Generic Object Detection (GOD). One of the most popular directions in computer vision is generic object detection, e.g., V. Sharma et al. [10], S. Rani et al. [11], CornerNet [12]. Remarkably, the generic objects can be salient or camouflaged. The camouflaged objects can be treated as difficult cases of generic objects. These methods have typically been successful in scenarios where the object foreground is obvious, however, experience has shown that they exhibit undesired performance degradation in the problems of concern in this paper and are not suitable for direct use in military high-level camouflage colors due to the complex environment and the high degree of inherent similarity between the foreground and background.

Camouflaged Object Detection. Traditional methods adopt hand-crafted features as a basis for discriminating camouflage objects, they focus on the texture area of the camouflage pattern or other important aspects thus presenting a variety of work to decipher the concealment. N.U. Bhajantri et al. [13] used a patch-level grayscale co-occurrence matrix to represent texture features and combined it with watershed segmentation and clustering strategies to successfully achieve the detection of camouflaged objects. L. Song et al. [14] designed an artificial texture descriptor including texture orientation, luminance, and entropy, which measures the reliability of foreground texture by the similarity of the weight structure of the features. F. Xue et al. [15] proposed a framework for quantitatively assessing the degree of difference between the object and the surrounding background by nonlinearly fusing multiple image features. Besides, Y. Chen et al. [16] proposed a scheme based on a convex optimization algorithm for mitigating the performance degradation of camouflage detection models in a more complex scene. However, those schemes based on hand-crafted features are effective only when the sample background is simple. They fail significantly in military high-level camouflage scenarios due to the high similarity between the object and the background. Because of the strong feature autonomous representation ability of the deep learning methods, we mainly discuss it in this paper.

In recent years, deep learning-based approaches such as convolutional neural networks (CNN), which specialize in image processing, have dominated computer vision research in several directions. Some scholars have applied it to the detection of plant and animal camouflage scenes. In 2019, Le et al. [17] proposed an end-to-end network consisting of classification and segmentation branches to cope with camouflaged object segmentation. The well-known SINet [18] was proposed by Fan et al. with two core components: the search module (SM), and the identification module (IM). Specifically, the classification stream is responsible for predicting the probability of objects included in the image while reinforcing the segmentation effect of the segmentation stream. In addition, Ge-Peng et al. [19] proposed an interesting edge-based reversible re-calibration network (ERRNet) by modeling visual perceptual behavior and cross-comparison between potential object regions and background. K. Wang et al. [20] designed a D2C-Net algorithm with Dual-Branch, Dual-Guidance, and Cross-Refine features to mimic the human visual mechanism of the two-stage detection process for concealed objects. Noting that severely constrained by public datasets, those methods mostly focus on semantic segmentation for binary-like operations, flora, and fauna rather than the 2D object detection, military camouflage (especially high-level camouflage masters) tasks which we focus on, and we are motivated by efforts to fill this research gap. A comprehensive review of camouflage segmentation is beyond the scope of this paper and the interested reader is referred to the wider literature [21] [22]. In recent closed-related works, He et al. [54] proposed the first fragile-supervised COD method by adopting scribble annotations as supervision, and achieved positive results. To cope with the visual challenge that the inherent similarity between the foreground object and the background environment usually brings deception to the algorithm and humans, Hu et al. [55] introduced a novel HitNet to refine the low-resolution representations with high-resolution features in an iterative feedback manner, thus avoiding the blurring of edges and boundaries caused by detail degradation. GP. Ji et al. [56] designed a novel deep framework that exploits object gradient supervision for camouflaged object detection, called DGNet. It decomposes the task into two connected branches composed of a context and a texture encoder, to skillfully realize the accurate perception of the foreground target hidden in the camouflage environment. Besides, FAP-net [57] is proposed for camouflaged object detection, which includes three precise core components, i.e., BGM module to provide boundary-enhanced features, MFAM module to obtain the aggregated feature representations, and CFPM module which can efficiently explore the cross-level correlations and transmit the valuable context information to enhance the performance of COD. Furthermore, Fan et al. [58] provided the latest survey on concerted scene understanding (CSU) to enable researchers to understand the global situation in the CSU field, including the current achievements and major challenges. Interested readers can read it in further detail to promote the vigorous development of advanced technologies and novel applications in this field.

To leverage features and improve performance, some detection or segmentation tasks also try to obtain more thorough feature cues by focusing on feature representation. For example, Lyu et al. [23] proposed a framework combining SOD and COD for camouflage targets for the first time and used an adaptive learning strategy and feature expression to effectively model the uncertainty of the model prediction. G. - P. Ji et al. [24] proposed a dispersion mining scheme by building a positioning and focus network to discover and eliminate the dispersion in the camouflage scene. Fan et al. [25] have developed a well-elaborated feature enhancement model including the subcomponents of neighbor connection decoder and group-level attention. In addition, to clearly explain the uncertainty in the camouflage scene,

Yang et al. [26] proposed a robust framework using the probabilistic representation model jointly with Transformers. Yet, their specific contribution to the field of military camouflage is continuously missing.

Different from the existing camouflaged object detection methods that rely on single target images of plants and animals, we propose a new scheme to achieve more accurate localization and boundary delineation for high-level military camouflaged objects. The obvious difference and advantages of the proposed strategy are that we incorporate global prior, relationship mining, and visual recovery of the bio-visual mechanism in the whole process, and we also design springboard selection to alleviate the sample domain imbalance problem. More promising results demonstrate that those mechanisms promote performance for MHCD.

3. Datasets

Datasets play a pivotal role in learning-based object detection methods. Unfortunately, to our knowledge, evaluation works on the existence of military camouflage object detection are rare and informal. To date still, no datasets designated for military camouflage and especially for high-level concealment grand-master are available. Thus, this has inspired our determination to promote this work. Domain-specific datasets are necessary because the characteristics of the objects present large gaps between different domains. Note that popular datasets for camouflage detection tasks [27] [28] [29] are mostly applicable only to the semantic segmentation of plants and animals. The closest to our work is [30], but it is also only for generic military objects (where the object is obvious) and not for high-level camouflage.

To construct the dataset, the images in MHCD2022 were collected from the website just as in [30]. To obtain a sample of real camouflage, we crawl the camouflage images immersed in the military context from the web search engine by setting multiple sets of keywords. In addition, we also collected several video clips and preserved the continuous image frames with 27 FPS. In total, the dataset includes five categories of person, aeroplane, military vehicle, warship, and tank, whose high-level camouflage involves a variety of realistic scenarios such as jungle, desert, snow, town, and ocean. Then, as a supplement, we also fused a small-scale dataset of 1000 single-person camouflages for semantic segmentation [2] and relabeled the bounding box labels. It is certainly suitable for the evaluation of advanced camouflage object detection methods for the military. It will continue to be extended in the future (including categories, diversity, scene attributes such as smoke, etc.).

After the sample collection was completed, data cleansing and manual filtering were specifically employed to eliminate unsatisfactory images. Images with serious visual problems caused by motion blur, low resolution, no military camouflage objects, high duplication, etc. were required to be removed before labeling. Finally, 3000 military high-level camouflage images were retained to ensure the appropriate quality of MHCD. Subsequently, each image is carefully labeled with instance categories and tightly wrapped rectangular boxes. Our dataset is available from the web after acceptance. At this stage of our work, we aim to achieve a precise attack on valuable military objects with low-cost data using lethal weapons. Due to the peculiarity and damage scope of the weapons, the requirement for a more accurate object contour is not so strong, and the bounding box can meet the task requirements to a certain extent. An important and more accurate mask annotation with a large amount of manpower consumption is also being made in full swing, to provide all-around data support in the future and further promote the application of scholars in a large-scale actual war. Our dataset is available on the website https://github.com/liumaozhen-lmz/Military-Camouflage-MHCD2022.git.

4. MHNet

In this section, we first illustrate the proposed overall architecture of MHNet as shown in Fig. 2. The four key components are respectively elaborated in detail.

4.1. Overall Architecture

In general, the human visual perception system intuitively receives subject information from the global visual field (which we call subject-aware prior), and the effective support related to subject perception lies in the learning and construction of global depth information by the model. In particular, VGG16 (after removing the fully connected layer) is used as the backbone to extract multi-depth features, denoted as F_i with i = 1,2,...,5. Our network is constructed based on the following considerations. Recent work [2] shows that the deeper output of the encoder contains a larger volume of semantic information, while the F_1 and F_2 layers in VGG mainly extract texture, color, and boundary details, which are crucial for camouflage detection. Therefore, we employ F_1, F_2 features used to represent the subject prior. Inspired by [17], we use F_3 as a bifurcation point. The route flow indicates that the PRM phase consists of F_4^2 and F_5^2 is used to integrate the potential relationship of the instance part with an object. The route flow indicates the CR/FCS module consisting of F_4^1 and F_5^1 is used to reverse the full characterization of the binding relationship, which is consistent with the observation that human minds make decisions when in the dilemma of military high-level camouflage. That is, the human first roughly perceives the object location from the global view, then searches for relevant *parts* (e.g. foot, shoulder) to form the whole object, and finally adds the detail cues (e.g, legs, connecting part of shoulder and body) lost by occlusion or texture drowning. To suppress the sample imbalance problem of the model (i.e., learning from hard samples), we propose the SS strategy (parallelogram region on light blue background in Fig. 2) at this stage. Ultimately, a fully informative and robust instance-level feature representation is generated for classification and coordinate localization.

4.2. Subject Perception Gathering

To provide a comprehensive and differentiated representation, we construct global perceptual ("abrupt" or inconsistent) features using shallow layers that retain richer detail information (e.g., texture, boundaries, color) than the deep layers which are cohesive with abstract semantics. Those well-endowed low-level representations are seen as the first level of augmented roles of the backbone stream (consisting of F_4^2 and F_5^2), which are fed to the PRM module, the backbone stream, at the output side by two 1×1 Conv weight controls to depict fine-grained information, respectively, which corresponds to the first global perception stage of the human visual mechanism. However, useful cues and interfering signals such as lines, background obstacles, redundant high-frequency signals, etc. simultaneously fill the low-level space. In addition, the down-sampling operation further introduces coarse cue loss in the deeper layers. In general, the direct use of the underlying representation to construct a geographic contour map-like approach may cause model oscillations and inconsistencies. The concept of SPG is illustrated in the pink parallelogram region at the bottom of Fig. 2.

To deal with those traps, here, with $F_1 \epsilon R^{H \times W \times 64}$ and $F_2 \epsilon R^{H \times W \times 256}$ as input, we introduce selective weighted attention aimed at adap-

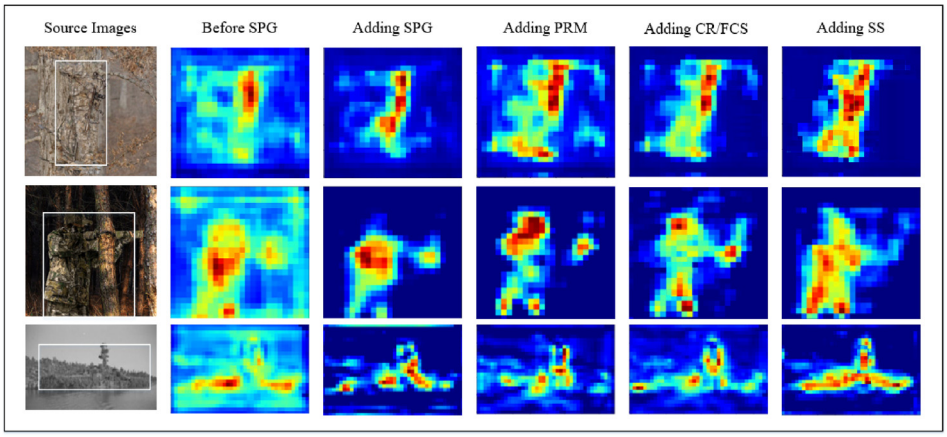

Fig. 2. The overall framework of our MHNet, which consists of four main components: Subject Perception Gathering (SPG), Part-object relationships mining (PRM), Concept recovery/feature clue supplement (CR/FCS), and Springboard Selection (SS). More details are described in Section 3.2∼Section 3.5.

tively extracting mutual representations before coalescence. Next, a 1 × 1 convolution computation with 32 filters is performed to reduce the number of parameters (compression in channel dimensions), have:

$$F'_1 = g^r_1(F_1), F'_2 = g^r_2(F_2) \tag{1}$$

Where $g^r_1(\cdot)$ stands for convolutional computation and r refers to the number of filters, r = 32. Next, we adopt a weighted regulator w^d_i to learn the level of attention to different branches, i = 1, 2. Formally, mathematics is defined as follows:

$$w^d = g^r\left(F'_1 \otimes F'_2\right) \tag{2}$$

Where $\otimes$ denotes element-wise multiplication. Note that our motivation is to employ this mutual weighting strategy, which can be viewed as linear attention, to suppress shallow nonrelevant noise and to give a coarse "obviousness" to the foreground. We find that it raises the problem of the disappearance of effective perceptual details, and we concatenate two characteristics that noise suppression and coarse perception based on the residual learning block, which can be expressed as:

$$\widetilde{F} = g^r\left(Cat\left(g^r\left(F'_1 \bigoplus w^d\right), g^r\left(F'_2\right) \bigoplus w^d\right)\right) \tag{3}$$

where Cat(·) refers to the concatenation operation along the channel dimension. Then, a classical Gaussian convolution operation is performed, which can be expressed as:

$$\widetilde{F}_G = GConv\left(\sigma\left(\widetilde{F}\right), k\right) \tag{4}$$

where GConv(·) denotes the Gaussian convolution operation with kernel k = 32 and zero bias. σ denotes the sigmoid function. This can provide the effect of Gaussian blurring to obtain transition camouflage maps with more comprehensive boundary information. The next "abrupt" geographic contour map is specifically generated with the efficient FPN [31]. Initial subject-aware prior associated with high-level military camouflage is suitable for feature optimization of mainstream and PRM streams. Intuitively, the FPN has two paths: one bottom-up path and one top-down path, as shown in the feature blocks drawn in dark blue and brown below Fig. 2. The bottom-up path is identical to the classical back-propagation network. We use only a three-layer structure by removing the last two layers and noting the original conv1, conv2, and conv3 as $\{C_1, C_2, C_3\}$. Each stage has a different scale due to the difference in the respective resolution of each stage. Then, a top-down propagation path is cast responsible for placing semantically stronger high-level representations into the lower layers of the network. In particular, P_3 is obtained by passing C_3 through a 1 × 1 convolution block. Each stage of features from high-level P_i with 2 times of upsampling, followed by summation with features from C_{i-1} through a 1 × 1 convolution block, corresponding to the feature maps noted as $\{P_1, P_2, P_3\}$ respectively. Compared to $\{C_1, C_2, C_3\}$, the latter has richer and different levels of abstract semantics and is all aware of advanced concealed styles. Finally, we perform upsampling and downsampling operations on P_1 and P_3 respectively to unify the scales and concatenate $\{P_1, P_2, P_3\}$ along the channel direction to generate a clear stereo depth-aware map (subject-aware prior) by a 1 × 1 convolution block.

4.3. Part-object relationships mining

Motivated by the successful exploration of the problem of incomplete object semantics in part-object relationships in [9], we decided to use CapsNet to perform the PRM sub-network based on the visual knowledge learned in the previous first level of augmentation, as the second level of feature enhancement. Specifically, a mirror CapsNet is included to capture the visual cues closely related to the object on the part. As the branch consists of F^2_4 and F^2_5 in Fig. 2, features from F^2_5 are efficiently transformed into capsule feature maps via Primary Capsule (PrimaryCaps). We introduce here a Convolutional Capsule (ConvCaps) layer and a Deconvolutional Capsule (DeconvCaps) layer for capsule routing, which generates a demanded mirror CapsNet via EM routing [32]. Thus, the part-object relations of the high-level military camouflage maps are mined. It noted that capsule embeddings are implemented upsampling for routing in the DeconvCaps stage, whose output contains part-object relations for capsule feature maps. Moreover, exhaustive details about PrimaryCaps and ConvCaps were able to be found in [33].

4.4. Concept recovery/feature clue supplement

As shown in Fig. 3, the biological mechanism of experience will strive to actively recover those precious features of the blocked parts after finding the objective rough location and relevant parts, i.e., reverse the exploration procedure. The details are elaborated in the green background area of Fig. 2. After the SPG, PRM, and SS modules have done their duty, it is necessary that the high-level camouflage map obtained in the mainstream be further refined in terms of the missing parts and the main details. Reverse attention [34] is an efficient and straightforward approach that stands out among many proposed works. It has the characteristic of pro-

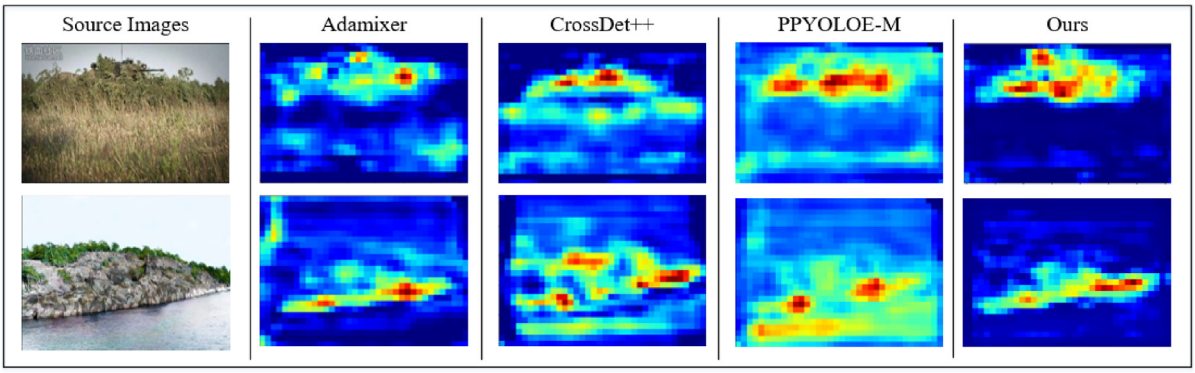

Fig. 3. Schematic diagram of feature omission with the general detector.

moting the module to discover complementary visual regions by clearing the predicted regions in the heat map. However, considering directly adopting the obtained weight map by reverse attention is not appropriate due to its containing redundant background noise, which will cause feature confusion and performance degradation at the decision end. To avoid the problem, we additionally fuse the boundary features from the previous module to make the model pay attention only to the missed feature cues inside the boundary constraints for learning.

Specifically, the ROI output from the deep layer with RPN is first computed by a 1 × 1 convolution to generate the masked feature map M^t. In this branch, the reverse attention map is obtained by reverse operation:

$$\alpha_C^t = 1 - Sigmoid(M^t) \tag{5}$$

where C refers to the channel. As a parallel branch, the information constraint f_C^t from F_5^1 with rich boundary features, and α_C^t will be multiplied with the encoded feature ROI^* in an element-wise manner for the recovery of missing body features, whose process can be described as:

$$x_C^t = Conv_{3\times3}\left(\alpha_C^t \times f_C^t \times ROI_C^*\right) \tag{6}$$

Where $Conv_{3\times3}$ denotes a 3 × 3 convolution block with 512 filters. Interestingly, it is noted that masked map M^t is a feature map and not a binary mask that often appears in segmentation. Because the feature map of ROI output after attention module has clearer foreground contour and shape, we define this fine example map as an example-level mask indicator M^t to guide the inner boundary constraints. At the same time, in the information constraint branch at the top of the CR/FCS module, the local area feature map of the corresponding instance with more details and texture information before attention is used as the outer boundary constraint to jointly define the search area of reverse attention in the CR/FCS module. Also, consider that the process suppresses attention to predicted regions and enhances the exploration of non-predicted regions. Predicted region features are changed and then perform selective weighted attention with previously pure ROI for mutual information compensation (described in subSection 4.2).

4.5. Springboard Selection

We know that the inevitable sample classification imbalance problem continuously threatens the efficiency of the model. Samples with different levels of discrimination difficulty in the data source (e.g., easy samples, hard samples) take a toll on the model's performance in accommodating both. The learning of the network is highly susceptible to slipping to easy-discrimination samples (e.g., the aeroplane in the seventh image of Fig. 1) while failing in the hard samples (e.g., the warship in the second image of Fig. 1). This is often studied as a separate direction, which we novelly translate to military high-level camouflage detection in this paper, and propose a springboard learning strategy to alleviate the problem.

From a probabilistic point of view, similar to the domain adaptive bottleneck [35], we consider that the general paradigm of the existing model is violently adaptive to all samples and lacks guidance. Notice that the classification imbalance in this task can be treated as a severe inconsistency in their data distribution. We define hard-discriminated samples and easy samples as x_h, x_e, respectively, and their data distributions are denoted as P_h and P_e. In fact, the camouflage detection problem can be treated as seeking the posterior $P(C, B|I)$, where B is the bounding box, $C \in \{1, \ldots, K\}$ is the class of the object and K is the total number of predefined classes (K = 5), and I is the image representation. For convenience, the joint data distribution of the samples used for training is denoted as P(C,B,I). Then, the joint distribution of hard samples is denoted as $P_h(C, B, I)$ and the easy samples as $P_e(C, B, I)$, with $P_h(C, B, I) \neq P_e(C, B, I)$. For the image-level balance, based on the Bayesian formulation the above joint distribution can be decomposed into the Eqn. (7).

$$P(C, B, I) = P(C, B|I) \cdot P(I) \tag{7}$$

Similar to [35] in making the covariate shift assumption, the conditional probability $P(C, B|I)$ is identical for hard and easy samples. In other words, we expect that the decision results can be maintained without caring about the effect of data distribution. Therefore, the oscillation or performance failure of this model is caused by marginal distribution P(I) differences to some extent. In the model, image representation I is the feature map of the F_4 layer output. To deal with the performance failure caused by extreme samples, we need to pay attention to enhancing the consistency of the marginal distribution of the two, so that the model can adapt to perform work on samples with lower variance, which can be expressed as:

$$P_h(I) \Longleftrightarrow P_e(I) \tag{8}$$

For instance-level balance, the joint probability distribution P(C,B,I) can be decomposed into Eqn. (9) based on the Bayesian formula:

$$P(C, B, I) = P(C|B, I) \cdot P(B, I) \tag{9}$$

We make the covariate shift assumption that the conditional probability $P(C|B, I)$ is identical for a sample of two degrees of discrimination. Then, this model incompatibility is caused by the difference in marginal distribution P(B,I). In other words, we expect the decision result to focus on the judgment of the correct instance margin (i.e., endowed with the same category label) regardless of the

degree of annihilation of the foreground by the background. Also, we naturally pay attention to the consistency of the distribution of their instance-level representations, which can be expressed as $P_h(B,I) \Longleftrightarrow P_e(B,I$. In our model, (B, I) denotes the features of the region in the feature map I which corresponds to the bounding box of the camouflaged instance ground truth.

For joint balance learning at the image level and instance level, for hard and easy samples, since $P(B,I) = P(B|I)P(I)$ and based on the assumption that $P(B|I)$ is identical and non-zero, it follows that:

$$P_h(I) = P_e(I) \Longleftrightarrow P_h(B,I) = P_e(B,I) \tag{10}$$

That is, the difference in the image-level feature representations should also be similar to instance-level representations if they converge to zero for both samples. The key to considering the problem should be to find ways to transform the easy samples through the encoder to conform to the distribution of the hard samples (i.e., $P_e(C,B,I) \Rightarrow P_h(C,B,I)$), forcing the model backend to receive and achieve uniform processing of the hard feature representations instead of processing the two with large differences simultaneously. However, the dynamic selection of intermediate springboards is necessary due to the variation of differences between them leading to a more difficult direct transformation. Without bells and whistles, we propose a SS strategy and a dynamic encoding activation method to alleviate the existence of gaps. We first take the sample feature to the mean of each batch size used for training as the initial springboard guide map (e.g., the upper gray block in the light blue background region of Fig. 2, with the data distribution noted as $P_\vartheta(C,B,I)$). Afterward, a subtract operation is performed and the parameters are compressed by a 1×1 convolutional block to obtain the feature map with the size of $152 \times 152 \times 256$. After that, two 1D features with the size of $1 \times 1 \times 256$ are obtained by global max-pooling and global average-pooling, respectively, due to the dual pooling having stronger representation power. Then, the two features are processed by a shared fully-connected layer(shared multi-layer perceptron, MLP) with the number of intermediate neurons being C/r(r is the compression coefficient as a parameter, C is the channel number of the feature, and r is 8 in this work, that is, 32) to finally output a dynamically encoded activation map, which preserves the fine-grained differences between the front-end features of each input map and the springboard. Notice that the significant differences between the two samples are decomposed into multiple gaps for processing instead of just one, i.e., $P_e(C,B,I) \Rightarrow P_\vartheta(C,B,I)_1 \Rightarrow P_\vartheta(C,B,I)_2 \Rightarrow \ldots \Rightarrow P_\vartheta(C,B,I)_m \Rightarrow P_h(C,B,I)$, m denotes the number of neurons in the fully connected intermediate layer. This is consistent with the idea of mathematical approximation. We regard this significant difference in domain distribution as a combination of multiple sub-gaps and use MLP to decode it to obtain modulation information. The transformation weighting guide $g \in R^{1\times1\times512}$ is generated by adding the results obtained from the shared fully-connected layer and then using the Sigmoid activation function to obtain the weight (between 0 and 1) of each channel in the feature map related to the difference in domain distribution to encode and activate the feature map channels of the backbone stream so that it maps to the data distribution of the biased hard samples. In detail, objects with a higher level of concealment will receive a smaller activation encoding while clear objects are activated more strongly. In this way, the back-end processing of the model tends to be solid benefiting from the transformation of the front encoding end. Notice that its different from SENet, which wants to enhance important features and weaken unimportant features, i.e. attention. The principle of our structure is to take the difference (or degree) of domain distribution as the input, which is similar to imposing punishment or disturbance to make the domain distribution of easy samples (with the large difference in domain distribution) complicated. However, when the distribution difference is small (that is, the input is a hard sample), this disturbance is relatively weakened and has little influence on the domain distribution, so that the subsequent network can focus on the difficult sample, thus making the extracted features more directional. Experimental results and visualization results prove the effectiveness of our method.

5. Experiments

5.1. Datasets and Implementation Details

We evaluate the performance of the proposed method on the military camouflage dataset MHCD. Our whole dataset is divided into a training set and a test set according to a 4-to-1 ratio. 2400 images are expanded to 12,000 images by classical image reversal, rotation, offset, and mosaic enhancement algorithms. Thus, 12,000 images are finally used for training and the remaining 600 images are used for testing. The specified categories and the number of instances are presented in Table 1. In addition, other features are summarized and compared with current popular object detection datasets (VOC2012 [36], COCO [37], and MOD [30]). For the density attribute regarding the number of boxes per image, the statistics are shown in Fig. 4.

We perform the proposed model under the Tensorflow framework. The deep learning environment is built on the Ubuntu v16.04 system, and all experiments are trained and evaluated on RTX3090 Ti 2GPU with 24 GB memory. Otherwise, we perform the training and testing procedure on images with a single scale and utilize Faster R-CNN as a baseline. Specifically, the input image is fixed to a size of 608 $\times$ 608. We use VGG16 as the feature extractor backbone and RoI Align as the ROI feature extractor. The weights of the backbone of the network are pre-trained on ImageNet. The batch size is set to 16 and the weight decay is 0.0001. The classical SGD optimizer with 0.9 momentum is used. Following the common evaluation metrics [30], we use mean Average Precision (mAP) to quantitatively evaluate the effectiveness of the network. Moreover, the network was trained for 30 K, 5 K, and 3 K iterations with a learning rate of 0.001, 0.0001, and 0.00001, respectively.

5.2. Comparisons with state-of-the-art methods

In Table 2, we compare the proposed MHNet with the state-of-the-art detectors under identical conditions. From the overall perspective, our method achieves robust results and can detect more concealed objects. First, MHNet achieves 56.76%, 38.16%, and 36.89% mAP with different threshold settings, respectively, and it presents 0.94%, 1.65%, and 0.87% mAP improvement compared to the Decoupled R-CNN method (i.e., 55.82% vs. 56.76%, 36.51% vs. 38.16%, 36.02% vs. 36.89%). In addition, our model is strongly competitive with AdaMixer at an IOU threshold setting of 0.5 close to which the latter performs best, AdaMixer reaching 56.83% of mAP with the ResNet-101 backbone. However, it is inferior to our MHNet in both the mAP@.75 and mAP@[.5,.95] metrics, which have the best results and present a clear advantage over recent methods. Notice that although our MHNet is second only to the

Table 1
The number of overall instances and categories in the dataset.

Categories	Number of overall instances
Person	11463
Tank	1627
Aeroplane	1139
Military Vehicle	915
Warship	628

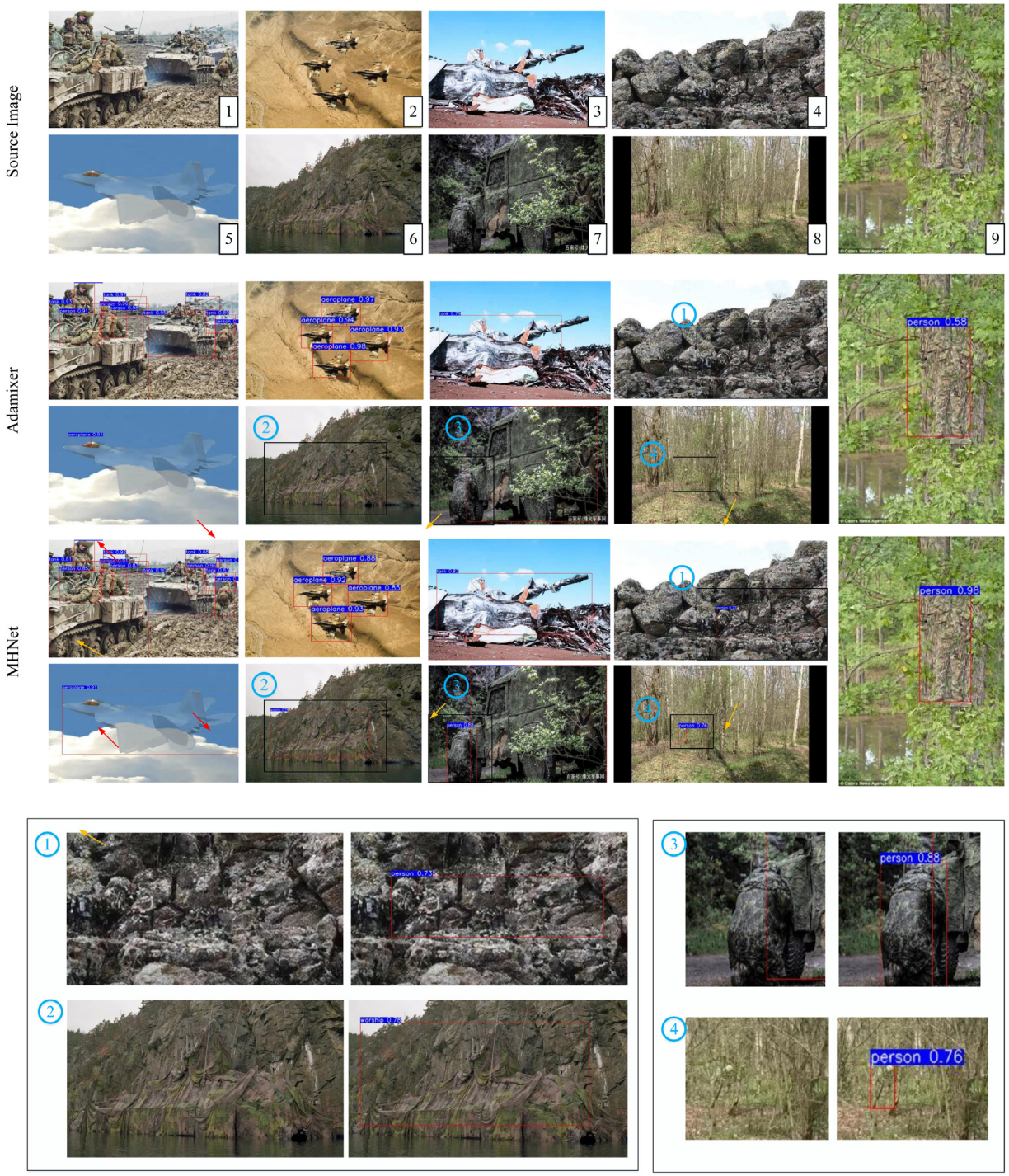

Fig. 4. Number of boxes per image after data augmentation.

AdaMixer algorithm in the "aeroplane", and "military vehicle" categories, this may be due to the relatively low level of camouflage of the relevant samples in the dataset. Also, as an important consensus [39], mAP@[.5,.95] is the most comprehensive rigorous metric to evaluate the overall performance of the model, because it is the average of mAP values corresponding to different thresholds. Although the method proposed in this work is not as good as AdaMixer in mAP@0.5, which is only 0.07% low, it maintains advantages in mAP@.75 and comprehensive average, which can prove the effectiveness of our method. Importantly, one of the highlights of our approach is its advantage in detecting objects with a high degree of camouflage. In other words, when the degree of camouflage of "aeroplanes" and "military vehicles" becomes high, the advantage of our model will be released as well. As shown in Fig. 5 for each group, the fifth image of a near-cloaked "aeroplanes" and the seventh image of "military vehicles" with a high degree of annihilation are both more accurately located by our method.

More, in Table 2, MHNet presents 0.98%, 2.03%, and 1.07% increases relative to ERFNet on the metrics of mAP, mAP@.75, and mAP@[.5,.95], respectively. Interestingly, compared to the PPYOLOE-M and CrossDet++ methods, our method has optimistic results on the five instance categories while maintaining the

Table 2
Quantitative comparison of MHNet with state-of-the-art object detection methods on the MHCD2022 dataset. The best results are highlighted with **bold** font. The second best results are underlined. The mAP and mAP@.75 indicate IOU thresholds of 0.5 and 0.75, respectively. mAP@[.5,.95] refers to mAP averaged for $IOU \in [0.5 : 0.05 : 0.95]$.

Methods	person	aeroplane	military vehicle	warship	tank	mAP	mAP@.75	mAP@[.5,.95]
SSD [38]	31.85	68.98	38.59	39.94	60.65	48.00	26.07	25.32
Faster R-CNN [39]	66.24	63.91	33.61	40.17	63.86	53.56	33.53	32.03
DETR [40]	65.72	67.04	39.73	45.53	62.29	56.58	36.22	34.71
FCOS [41]	63.15	65.21	32.78	42.32	60.19	52.73	33.91	32.65
Retinanet [42]	64.40	63.68	34.73	41.56	61.16	53.12	35.28	33.97
Cascade R-CNN [43]	66.82	64.31	31.08	43.57	62.24	53.60	38.03	36.28
Sparse R-CNN [44]	62.39	62.57	30.56	42.34	61.78	51.93	32.47	32.11
Decoupled R-CNN [45]	65.74	67.25	38.91	46.22	64.29	55.82	36.51	36.02
ERFNet [46]	64.32	66.71	39.17	45.51	63.58	55.78	36.13	35.82
PPYOLOE-M [47]	65.26	63.81	38.65	46.49	63.41	55.98	36.70	35.91
AdaMixer [48]	64.55	**69.79**	39.67	**47.80**	64.07	**56.83**	37.92	36.53
CrossDet++ [49]	66.98	64.79	39.71	46.59	64.10	56.42	37.84	36.62
MHNet(Ours)	**67.37**	65.22	**40.03**	46.87	**64.31**	56.76	**38.16**	**36.89**

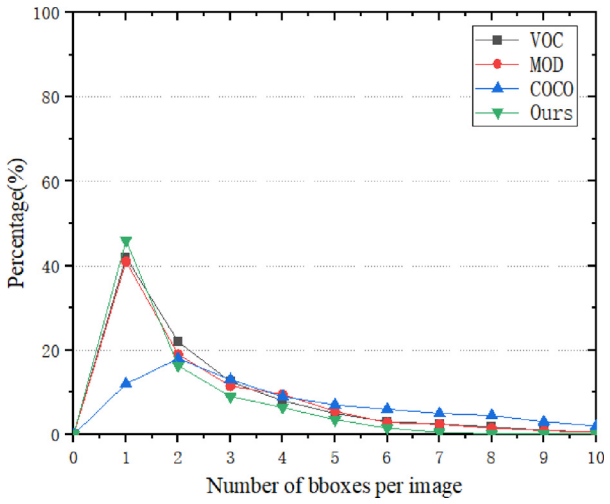

Fig. 5. Qualitative comparison results of MHNet with Adamixer. A score threshold of 0.5 is used in visualization.

advantage on mAP, mAP@.75, and mAP@[.5,.95]. Therefore, our method has credibility.

To further demonstrate this factor and to emphasize the effectiveness of MHNet on high-level concealed objects, the qualitative experimental results are specifically printed in Fig. 5. It can see that the most threatening AdaMixer method performs better in scenes with lower levels of camouflage, such as the first and second images in each group, and MHNet can effectively distinguish them with lower confidence scores (indicated by the red arrow). However, our method has a much better discovery ability and robust performance for high-level concealed objects. In detail, AdaMixer loses 'person' and 'warship' in the 4th and 6th images, respectively. In addition, the 'person' in the lower left corner of the 7th image and the 8th jungle are ignored. More, although AdaMixer perceives the objects in both 3, 5, and 9, our MHNet has more desirable confidence and localization power (indicated by the yellow arrow). They are both difficult to detect correctly by the human eye without a close look. Thus, our MHNet appears to be more effective in military high-level camouflage scenarios and has the potential to be extended to individual combat systems. In addition, the results of feature-level comparisons between several methods are also provided in Fig. 6. Obviously, our method has better feature representation capability.

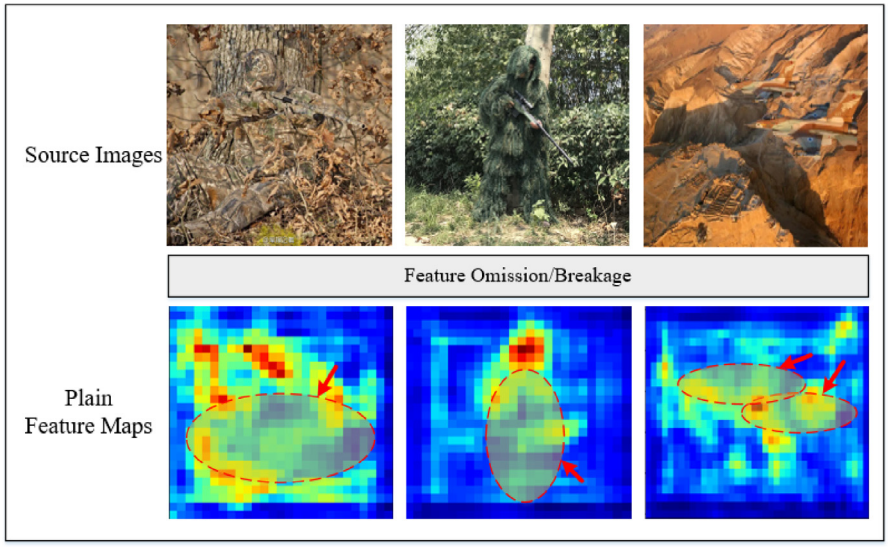

Fig. 6. Experimental results of feature-level comparisons for different methods.

5.3. Ablation Studies

As previously mentioned, the proposed approach contributes four interesting components, i.e., Subject Perception Gathering (SPG), Part-object relationships mining (PRM), Concept recovery/feature clue supplement (CR/FCS), and Springboard Selection (SS). To analyze the importance and effectiveness of the respective modules, an ablation study was conducted in Table 3.

Impact of the SPG module. We first evaluate the impact of the designed SPG module by comparing it with the network baseline. As can be seen in Table 3, the proposed SPG significantly outperforms the baseline substantially on all three criteria, it has 54.73% mAP and improves by 1.17%. This clearly indicates its global representation capability and the efficiency of deep stereo features.

What is the effect if we continue to add PRMs? Based on the efforts of the SPG approach set as a first-level enhancement role, we evaluate the positive effect of the PRM being used as a second-level enhancement. As can be seen from the third row of the table, with the introduction of the PRM module, MHNet achieves notable gains over the "+SPG", appears to be optimized above 1.08%, and achieves 55.81%, 35.68%, and 33.42% discriminative accuracy, respectively. Moreover, it can be seen that using PRM before SPG is not as effective as using SPG-PRM, due to the lack of subject-aware prior making PRM unable to do precise relationship mining, and there is no rough guide to determine the foreground and background.

The CR/FCS module in the network. Here, we further investigate the contribution of CR/FCS. Intuitively, it has a 0.51% mAP value improvement from 55.81% to 56.32%. This proves that the module successfully recovers the interrupted desired features caused by the deep computation and the occlusion interference terms in the source images.

The validity of SS. As shown in the fifth row of Table 3, our proposed MHNet presents a consistent boost on all listed common metrics, and its phenomenon demonstrates that the supremacy of springboard selection enables the model to focus more on balanced learning and suppress the occurrence of inconsistent gaps in the sample domain distribution.

5.4. Qualitative Analysis

To further understand the effects of SPG, PRM, CR/FCS, and SS, we visualized and analyzed the features before and after them separately. As shown in Fig. 7, after the SPG the global subject features are activated to benefit the subsequent discrimination procedure (see the third column for the main feature perception), and the rough location of the concealed object is captured. After that, as shown in the fourth column, in the first line, the PRM further succeeds in activating the foot (part) feature expressions closely related to the instance (object). However, it can observe the mentioned undesired object interruption phenomenon due to multifactorial interference, and our proposed CR/FCS module can explore

Table 3
Ablation experiments of the proposed MHNet on the MHCD2022 dataset. The best results are highlighted with **bold** font. "✔-"refers to the experimental results of applying PRM before SPG.

SPG	PRM	CR/FCS	SS	mAP@.5	mAP@.75	mAP@[.5,.95]
				53.56	33.53	32.03
✔				54.73	33.79	32.86
	✔			54.39	33.72	33.32
		✔		53.98	33.97	33.18
			✔	53.96	33.93	33.07
✔	✔			55.81	35.68	33.42
✔-	✔-			55.69	35.40	33.37
	✔	✔		54.73	34.82	33.31
	✔		✔	54.69	34.77	33.26
		✔	✔	54.61	34.59	33.20
✔	✔	✔		56.32	36.03	34.71
	✔	✔	✔	55.87	37.21	34.06
✔	✔	✔	✔	**56.76**	**38.16**	**36.89**

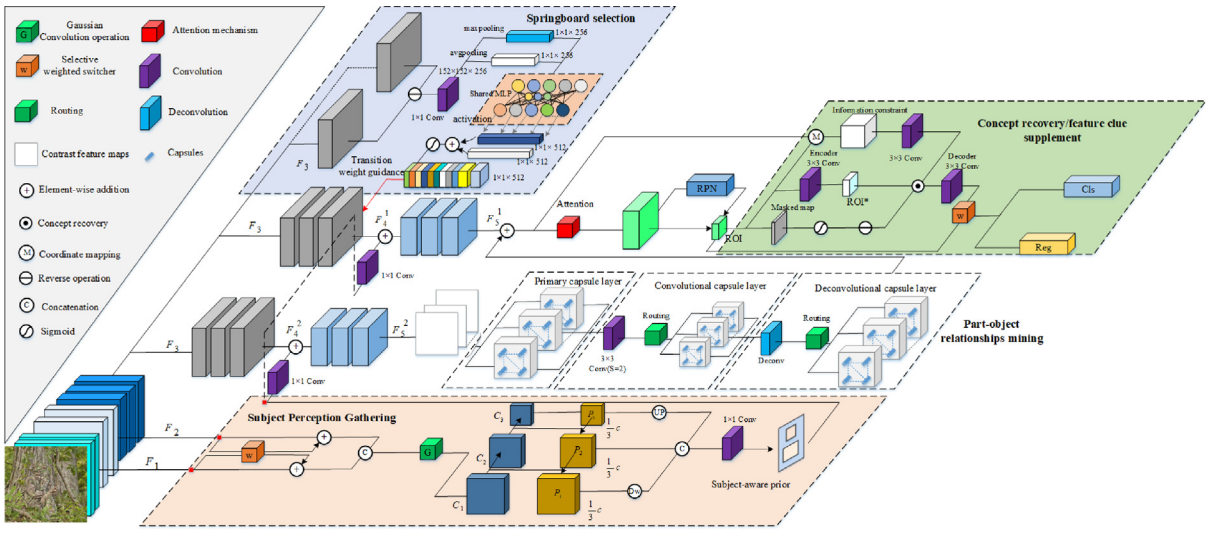

Fig. 7. Visualization and effect of the features before and after different core components.

Fig. 8. Selected video examples of object detection results on the MHCD2022 test set using the MHNet system. Our method detects objects of a wide range of scales and aspect ratios. Each output box is associated with a category label and a softmax score in [0,1]. A score threshold of 0.5 is used to display these images.

the missed feature expressions to some extent as shown in the fifth column in Fig. 7, where the human leg features are specifically activated to recover such that the overall object is more visible. Due to the favorable knowledge learning of camouflaged objects, the SS module improves the model's feature representation of camouflaged targets, which is beneficial for detection. Of course, both the second and third rows of results have a similar effect, and the bounding box associated with the object in the source image is also provided. More detection results are shown in Fig. 8.

6. Conclusion and Future work

This paper investigates for the first time high-level military concealed object detection. Specifically, we have provided the new challenging and carefully labeled MHCD2022 benchmark and developed a novel but efficient end-to-end perception and identification framework (i.e. MHNet) based on human visual mechanisms. Our MHNet achieves robust and superior results when tackling objects with a high degree of concealment compared to other works that exist. The above contributions have provided the military community with ideas for designing new models for the MHCD task. In the future, we plan to expand the MHCD2022 dataset to provide a variety of categories including missiles (type level), amphibious tanks, soldiers as well as officers, civilian vehicles, civilian facilities, etc, pixel-level labels(i.e., mask annotation) for semantic segmentation are about to be produced. Interestingly, in the near future, it will carry out more in-depth research with multiple segmented data sets such as COD10K. In addition, we also plan to work on fusing visible and infrared approaches while designing more advanced networks based on great frameworks and making an effort to focus on deep learning interpretive work.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the Natural Science Foundation of Heilongjiang Province of China under Grant No. LH2021F026; the Fundamental Research Funds for Central Universities under Grant No.HIT.NSRIF202243.

References

[1] N.E. Scott-Samuel, R. Baddeley, C.E. Palmer, I.C. Cuthill. Dazzle camouflage affects speed perception, in PLoS One, 2011, pp. 6.
[2] Z. YunFei, Z. Xiongwei, F. Wang, C. Tiieyong, S. Meng, W. Xiaobing, Detection of People With Camouflage Pattern Via Dense Deconvolution Network, in IEEE Signal Processing Letters, 2018, PP. 1–1. DOI: 10.1109/LSP.2018.2825959.
[3] M. Stevens, I.C. Cuthill, A.M.M. Windsor, H.J. Walker, Disruptive contrast in animal camouflage, in PoRS, Biological Sciences (2006) 2433–2438.
[4] M. Dimitrova, N. Stobbe, H.M. Schaefer, S. Merilaita, Concealed by conspicuousness: distractive prey markings and backgrounds, in PoRSB, Biological Sciences (2009) 1905–1910.
[5] S. Astapov, J.-S. Preden, J. Ehala, A. Riid, Object detection for military surveillance using distributed multimodal smart sensors, in: 2014 19th international conference on digital signal processing, IEEE, 2014, pp. 366–371.
[6] Z. Yang, W. Yu, P. Liang, H. Guo, L. Xia, F. Zhang, Y. Ma, J. Ma, Deep transfer learning for military object recognition under small training set condition, Neural Computing and Applications 31 (10) (2019) 6469–6478.
[7] L. Tang, B. Li, S. Kuang, et al., Re-thinking the relations in co-saliency detection, in IEEE Transactions on Circuits and Systems for Video Technology, 2022.
[8] Z. Yao, L. Wang, Boundary Information Progressive Guidance Network for Salient Object Detection, in IEEE Transactions on Multimedia, 2021, 24: 4236–4249.
[9] Y. Liu, D. Zhang, Q. Zhang, et al., Integrating part-object relationship and contrast for camouflaged object detection, in IEEE Transactions on Information Forensics and Security, 2021, 16: 5154–5166.
[10] H. Law and J. Deng, Cornernet: Detecting objects as paired keypoints, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 734–750.
[11] Y. Lyu, J. Zhang, Y. Dai, L. Aixuan, B. Liu, N. Barnes, D.-P. Fan, Simultaneously localize, segment and rank the camouflaged objects, in: IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[12] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, D.-P. Fan, Camouflaged object segmentation with distraction mining, in: IEEE Conf. Comput. Vis. Pattern Recog., 2021.
[13] N.U. Bhajantri, P. Nagabhushan, Camouflage defect identification: A novel approach, in: Proc. 9th Int. Conf. Inf. Technol., 2006, pp. 145–148.
[14] L. Song, W. Geng, A new camouflage texture evaluation method based on WSSIM and nature image features, in: Proc. Int. Conf. Multimedia Technol., 2010, pp. 1–4.
[15] F. Xue, C. Yong, S. Xu, H. Dong, Y. Luo, W. Jia, Camouflage performance analysis and evaluation framework based on features fusion, Multimedia Tools Appl. 75 (7) (2016) 4065–4082.
[16] Y. Pan, Y. Chen, Q. Fu, P. Zhang, X. Xu, Study on the camouflaged target detection method based on 3D convexity, Modern Appl. Sci 5 (4) (2011) 152–157.
[17] T.-N. Le, T.V. Nguyen, Z. Nie, M.-T. Tran, A. Sugimoto, Anabranch network for camouflaged object segmentation, in Comput, Vis. Image. Underst 184 (2019) 45–56.
[18] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, L. Shao, Camouflaged object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2020, pp. 2777–2787.
[19] G.-P. Ji, L. Zhu, M.C. Zhuge, K. Fu, Fast Camouflaged Object Detection via Edge-based Reversible Re-calibration Network, Pattern Recognition, Volume 123, 2022, 108414, ISSN 0031–3203.
[20] K. Wang, H. Bi, Y. Zhang, et al., D2C-Net: A Dual-Branch, Dual-Guidance and Cross-Refine Network for Camouflaged Object Detection, IEEE Trans. Ind. Electron. 69 (5) (2021) 5364–5374.
[21] H. Bi, C. Zhang, K. Wang, et al., Rethinking Camouflaged Object Detection: Models and Datasets, in IEEE Transactions on Circuits and Systems for Video Technology, 2021.
[22] C. Tianyou, X. Jin, H. Xiaoguang, Z. Guofeng, W. Shaojie, Boundary-guided network for camouflaged object detection, in Knowledge-Based Systems, Volume 248, 2022, 108901, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2022.108901.
[23] V. Sharma, R. N. Mir, Saliency guided faster-RCNN (SGFr-RCNN) model for object detection and recognition, in Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 5, 2022, Pages 1687–1699, ISSN 1319–1578, doi: 10.1016/j.jksuci.2019.09.012.
[24] S. Rani, D. Ghai, S. Kumar, Object detection and recognition using contour based edge detection and fast R-CNN, in Multimed Tools Appl, 2022, vol. 81, pp. 42183–42207. doi: 10.1007/s11042-021-11446-2.
[25] D.-P. Fan, G.-P. Ji, M.-M. Cheng, L. Shao, Concealed object detection, IEEE T. Pattern Anal. Mach. Intell. (2021).
[26] F. Yang, Q. Zhai, X. Li, R. Huang, A. Luo, H. Cheng, D.-P. Fan, Uncertainty-guided transformer reasoning for camouflaged object detection, in: Int. Conf. Comput. Vis., 2021.
[27] P. Skurowski, H. Abdulameer, J. Blaszczyk, T. Depta, A. Kornacki, and P. Koziel, Animal camouflage analysis: Chameleon database, in Unpublished Manuscript, vol. 2, no. 6, p. 7, 2018.
[28] T.-N. Le, T.V. Nguyen, Z. Nie, M.-T. Tran, A. Sugimoto, Anabranch network for camouflaged object segmentation, Comput. Vis. Image Understand. 184 (Jul. 2019) 45–56.
[29] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, L. Shao, Camouflaged object detection, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 2777–2787.
[30] Y. Xin, W. Jiahao, M. Bo, O. Yangtong, L. Longyao, MOD: Benchmark for Military Object Detection, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021.
[31] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, 2016, arXiv preprint arXiv:1612.03144.
[32] S. Sabour, N. Frosst, G. Hinton, Matrix capsules with em routing, in: Proc. Int. Conf. Learn. Represent, 2018, pp. 1–15.
[33] Y. Liu, Q. Zhang, D. Zhang, and J. Han, Employing deep part-object relationships for salient object detection, in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1232–1241.
[34] X. Xiuqi, Z. Mingyu, Y. Jinhao, C. Shuhan, H. Xuelong, Y. Yuequan, Boundary guidance network for camouflage object detection, in Image and Vision Computing, Volume 114, 2021, 104283, ISSN 0262-8856, https://doi.org/10.1016/j.imavis.2021.104283.
[35] Y. Chen, H. Wang, W. Li, et al., Scale-Aware Domain Adaptive Faster R-CNN, in Int J Comput Vis, 2021, vol. 129, 2223–2243. doi: 10.1007/s11263-021-01447-x.
[36] M. Everingham, S.A. Eslami, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, Int. J. Computer Vision 111 (1) (2015) 98–136.
[37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
[38] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37.
[39] S. Ren, K. He, R. Girshick, and J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in Advances in neural information processing systems, 2015, pp. 91–99.
[40] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, End-to-end object detection with transformers, In ECCV, 2020.
[41] T. Zhi, S. Chunhua, C. Hao, and H. Tong, FCOS: fully convolutional one-stage object detection, In ICCV, 2019.
[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, Focal loss for dense object detection, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
[43] Z. Cai and N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154–6162.
[44] S. Peize, Z. Rufeng, J. Yi, K. Tao, X. Chenfeng, Z. Wei, M. Tomizuka, L. Lei, Y. Zehuan, W. Changhu, and L. Ping, Sparse R-CNN: end-to-end object detection with learnable proposals, In CVPR, 2021.
[45] D. Wang, K. Shang, H. Wu, et al., Decoupled R-CNN: Sensitivity-Specific Detector for Higher Accurate Localization, in IEEE Transactions on Circuits and Systems for Video Technology, 2022.
[46] Qijin Wang, Yu. Shengyu Zhang, Guangcai Zhang Qian, Hongqiang Wang, Enhancing representation learning by exploiting effective receptive fields for object detection, Neurocomputing 481 (2022) 22–32, ISSN 0925–2312..
[47] X. Shangliang, W. Xinxin, L. Wenyu, C. Qinyao, C. Cheng, D. Kaipeng, W. Guanzhong, D. Qingqing, W. Shengyu, D. Yuning, et al., PP-YOLOE: An evolved version of YOLO, arXiv preprint arXiv:2203.16250, 2022.
[48] Z. Gao, L. Wang, B. Han, et al., AdaMixer: A Fast-Converging Query-Based Object Detector, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5364–5373.
[49] H. Qiu et al., CrossDet++: Growing Crossline Representation for Object Detection, IEEE Trans. Circuits Syst. Video Technol. 33 (3) (March 2023) 1093–1108, https://doi.org/10.1109/TCSVT.2022.3211734.
[50] L. Tang, J. Yuan, J. Ma, Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network[J], Inform. Fusion 82 (2022) 28–42.
[51] J. Ma, L. Tang, F. Fan, et al., SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer[J], IEEE/CAA J. Automatica Sinica 9 (7) (2022) 1200–1217.
[52] L. Tang, Y. Deng, Y. Ma, et al., SuperFusion: A versatile image registration and fusion network with semantic awareness[J], IEEE/CAA J. Automatica Sinica 9 (12) (2022) 2121–2137.
[53] Liu J, Fan X, Huang Z, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5802–5811.
[54] He R, Dong Q, Lin J, et al. Weakly-Supervised Camouflaged Object Detection with Scribble Annotations[J]. arXiv preprint arXiv:2207.14083, 2022.
[55] Hu X, Fan D P, Qin X, et al. High-resolution Iterative Feedback Network for Camouflaged Object Detection[J]. arXiv preprint arXiv:2203.11624, 2022.
[56] G.P. Ji, D.P. Fan, Y.C. Chou, et al., Deep Gradient Learning for Efficient Camouflaged Object Detection, Mach. Intell. Res. 20 (2023) 92–108, https://doi.org/10.1007/s11633-022-1365-9.
[57] T. Zhou, Y. Zhou, C. Gong, et al., Feature Aggregation and Propagation Network for Camouflaged Object Detection[J], IEEE Trans. Image Processing 31 (2022) 7036–7047.
[58] Fan, Deng-Ping, et al. Advances in Deep Concealed Scene Understanding. arXiv preprint arXiv:2304.11234 (2023).

Maozhen Liu was born in Jining, Shandong, China, in 1995. He is currently pursuing the Ph.D. degree in the Department of Astronautics, Harbin Institute of Technology. His main research interests include deep learning and object detection.

Xiaoguang Di was born in Heilongjiang, China. He received the M.S. and Ph.D. degrees in navigation, guidance and control from Northwestern Polytechnical University, China, in 1999 and 2004, respectively. He is currently a Professor with the Control and Simulation Center, Harbin Institute of Technology. His current research interests include real-time image restoration and enhancement, 2D and 3D object detection and recognition, deep learning, and SLAM. He is a member of the Chinese Association of Automation, the China Simulation Federation, and the Chinese Society of Astronautics.