Optical Engineering

OpticalEngineering.SPIEDigitalLibrary.org

Enhanced situation awareness through CNN-based deep multimodal image fusion

Shuo Liu
Huan Liu
Vijay John
Zheng Liu
Erik Blasch

SPIE.

Shuo Liu, Huan Liu, Vijay John, Zheng Liu, Erik Blasch, "Enhanced situation awareness through CNN-based deep multimodal image fusion," *Opt. Eng.* **59**(5), 053103 (2020), doi: 10.1117/1.OE.59.5.053103

Enhanced situation awareness through CNN-based deep multimodal image fusion

Shuo Liu,[a,†] Huan Liu,[a,b,†] Vijay John,[c] Zheng Liu,[a,*] and Erik Blasch[d]
[a]University of British Columbia Okanagan, School of Engineering, Kelowna, British Columbia, Canada
[b]China University of Geosciences, School of Automation, Wuhan, Hubei, China
[c]Toyota Technological Institute, Research Center of Smart Vehicles, Nagoya, Japan
[d]Air Force Research Laboratory, Rome, New York, United States

Abstract. Automated situation awareness (ASA) in a complex and dynamic setting is a challenging task. The accurate perception of environmental elements and events is critical for the successful completion of a mission. The key technology to implement ASA is target detection. However, in most situations, targets of interest that are at a distance are hard to identify due to the small size, complex background, and poor illumination conditions. Thus, multimodal (e.g., visible and thermal) imaging and fusion techniques are adopted to enhance the capability for situation awareness. A deep multimodal image fusion (DIF) framework is proposed to detect the target by fusing the complementary information from multimodal images with a deep convolutional neural network. The DIF is built and validated with the Military Sensing Information Analysis Center dataset. Extensive experiments were carried out to demonstrate the effectiveness and superiority of the proposed method in terms of both detection accuracy and computational efficiency. [DOI: 10.1117/1.OE.59.5.053103]

Keywords: situation awareness; image fusion; convolutional neural networks; target detection.

Paper 20200413 received Apr. 11, 2020; accepted for publication May 5, 2020; published online May 19, 2020.

1 Introduction

Automated exploitation is paramount in modern complex systems; it allows human operators to respond immediately and take actions to increase the survivability and security of the equipment, platforms, and forces.[1] In the tasks of target detection and scene perception, automated situation awareness and surveillance become prominent mechanisms to assist human operators with the extracted information, reliable evidence, and extended perception.

To address the challenges arising from the complex scenarios, multimodal image fusion techniques are often employed.[2–6] By fusing the complementary cross-spectrum information acquired through a multimodal imaging system, the target can be detected from a complex background or from a long distance. In our previous work,[7] multimodal images were fused with a shallow convolutional neural network (CNN) model, and a fast regions with convolutional neural network (R-CNN)[8] framework was adopted to detect the target. Even though the results are promising in comparison with unimodal imaging methods, there is still the room to improve the run-time as well as the accuracy for use in a practical application.

Actually, many challenges exist for automated applications, including target scale variations, environmental diversity, and real-time response requirements. In most scenarios, the scene is vast and expansive as illustrated with the examples on the left side of Fig. 1. The different distances from the imaging sensors to the target dramatically vary the scale of the target. The sample image on the right side of Fig. 1 demonstrates the complexity of a scenario. The target is hard to discriminate from the complex background from its (color and texture) appearance due to the camouflage of the target. Moreover, rocks and trees can also obscure the targets.

*Address all correspondence to Zheng Liu, E-mail: zheng.liu@ubc.ca
†Co-first author: These authors contributed equally to this work.

0091-3286/2020/$28.00 © 2020 SPIE

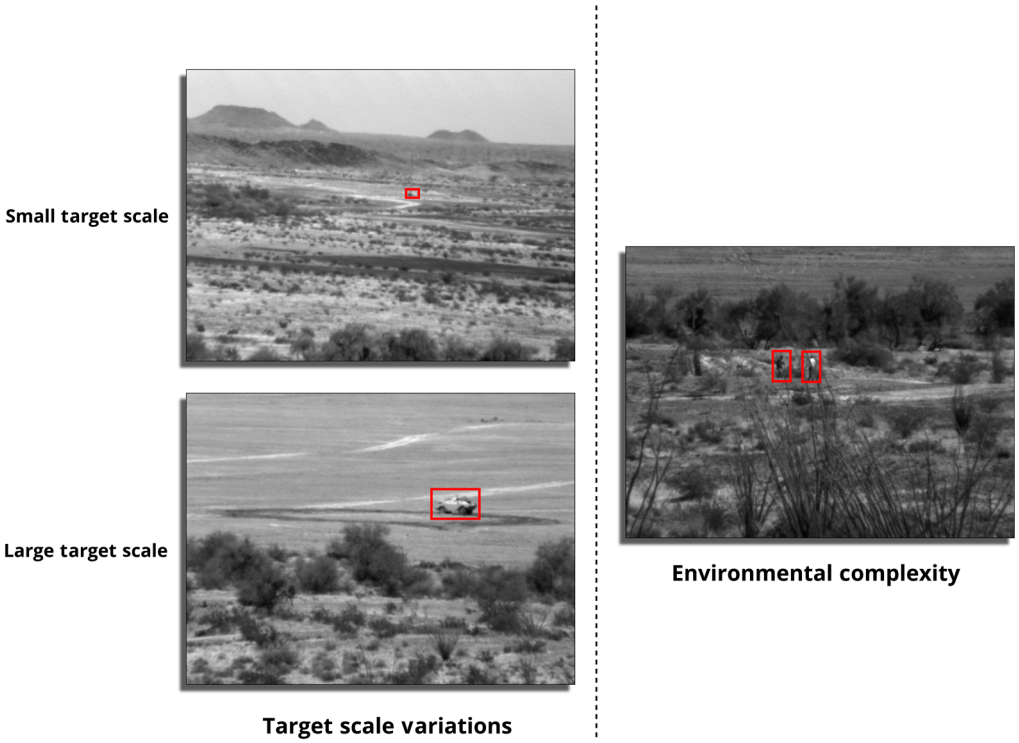

Fig. 1 Two sample images from the SENSIAC dataset illustrating complex situations.

These environmental factors will limit automated applications, especially for automatic target recognition (ATR). In addition, automated surveillance must have the capability to operate around-the-clock and provide immediate indications, warnings, and responses, thereby increasing requirements for robust and real-time performance.

Current research on automated surveillance is mainly focused on ATR applications, such as object classification, target tracking, identification, etc.,[9–12] rather than target detection, which is fundamental and important. A detailed review of the state-of-the-art methods is presented in Sec. 2.

In this study, we extended our previous fusion algorithm to improve its accuracy and efficiency for a robust performance and conducted a comprehensive analysis and extensive validating experiments. Specifically, a deeper CNN model was adopted to carry out deep feature extraction/fusion from multimodal images and perform target detection tasks. In addition, the handcrafted region proposal module, "selective search," [13] used in the previous work[7] was replaced with a more efficient module "region proposal network (RPN)."[14] Thus, the proposed deep image fusion (DIF) framework is a full end-to-end neural network, which can be optimized on a graphics processing unit (GPU) device. Moreover, a comprehensive analysis for complex scenarios was performed in the experimental section to show the effectiveness of the proposed framework. Figure 2 presents the overall architecture of the DIF framework, which consists of

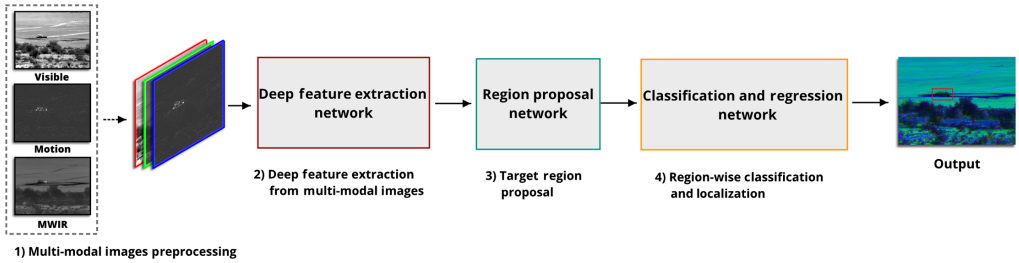

Fig. 2 Overall framework of the proposed DIF, including three major networks: (1) deep feature extraction network, (2) RPN, and (3) classification and regression network.

three main networks: deep feature extraction/fusion network, RPN, and classification and regression network.

In this paper, a DIF method that is capable of learning the complementary information from visible, thermal, and temporal images automatically for target detection is proposed. The implementation is in the form of an efficient end-to-end framework, which is based on the architecture of the CNN. The new framework integrates the deep feature extraction network, RPN, and classification and regression network to achieve a higher detection accuracy as well as computational efficiency. The performance is validated with the SENSIAC dataset in comparison with the state-of-the-art methods.

The rest of the paper is organized as follows. Section 2 reviews the state-of-the-art methods of the relevant work. The detailed description of the proposed method is presented in Sec. 3. Extensive experimental results are give in Sec. 4. Section 5 concludes this paper.

2 Review of Related Work

2.1 *Situation Awareness and Surveillance*

The capability of around-the-clock operations requires situation assessment, such as military surveillance for national defense. In general, a surveillance system comprises five key components:[15] target detection, tracking, classification, recognition, and identification. Among these components, target tracking aims to track one or multiple targets over time based on a given accurate location. Gundogdu et al.[16] proposed an ensemble tracking algorithm that is able to switch different correlators according to the current target appearance. Even though they achieved promising accuracy, there was still room for further improvement on computational efficiency. To this end, Demir and Cetin[17] implemented an efficient tracker by leveraging the codifference matrix. In addition to tracking, research is also focused on high-level tasks, such as ATR. A series of studies proposed a shape generative model-based general system that supports recognition, segmentation, and pose estimation jointly.[9,12,18]

Target detection is a fundamental and important component of a surveillance system, especially in challenging conditions in military contextz. However, only a few reports for military target detection are publicly available to the best of our knowledge. Most recently, Millikan et al.[19] proposed an infrared (IR)-focused military target detector combining both image reconstruction and quadratic correlation techniques. But the accuracy is not sufficient enough to be applied in a real scenario. In fact, generic target detection is an active research field that aims to recognize and localize one or more objects from an image or a video clip. The last decade has witnessed a revolution in this field, from traditional methods to CNN-based methods. In the traditional approaches, the representative work is the deformable part models,[20] which follow the conventional paradigm of sliding window templates trained by the latent support vector machine (SVM) with the histogram of oriented gradients feature. For the CNN-based approaches, Sermanet et al.[21] first utilized CNN models in a sliding window fashion on the generic target detection task, where two CNNs are involved, e.g., one for classifying if a window contains a target and the other for predicting the corresponding bounding boxes. Subsequently, the dominant CNN-based target detection framework, R-CNN, was proposed.[22] This framework utilizes a pretrained CNN to extract features on the region of interests (ROIs) generated by selective search[13] and classifies them with class-specific linear SVMs. The significance of this work is replacing hand-engineered features with the features extracted by CNN. Moreover, the variants of R-CNN, such as SPP-Net[23] and fast R-CNN,[8] were proposed to solve the computational efficiency issue. Most recently, Ren et al.[14] created a framework called "faster R-CNN" in which the region proposal module is replaced by an RPN. The feature extraction of RPN is shared with fast R-CNN, so they can be trained jointly. This method greatly improves the accuracy and efficiency of the fast R-CNN algorithm. Moreover, Liu et al.[24] proposed a more efficient target detector, named "SSD," which removed the region proposal module and only utilized a single network to accomplish target detection. Again, a report on using CNN for military target detection is not available. In this study, our effort is to migrate the success from the generic target detection, i.e., faster R-CNN, with a deeper CNN (ResNet 101) to the military target detection challenge.

2.2 *Multimodal Image Fusion*

There still exist numerous challenges that need to be solved in surveillance system designs. First of all, the scale of target varies over a range. Specifically, the scenario is expansive and the target of interest may be extremely far from the surveillance devices and sensors. As a result, the scale of the targeted object captured through the image/video is rather small, and the target can not be easily detected. The second challenge is the complex environment of the scenario. Rocks and trees may obscure the target. Meanwhile, the targeted objects are likely to disguise themselves, so they can not easily be recognized by the surveillance system.

Multimodal image fusion techniques can offer an effective solution to such challenges.[25,26] The fusion operation will generate a composite image with complementary information from multimodal images acquired through a wider range of the electromagnetic spectrum. The high-level surveillance tasks will be carried out based on the fused outcomes. For example, Zheng and Blasch[27] improved the performance of vehicle identification and threat analysis via multimodal image fusion. In particular, IR/thermal image and visible image (VI) are widely adopted in multimodal imaging systems for military applications. The fusion operation can be implemented at pixel-, feature-, and decision-levels. Numerous works on pixel-level fusion has been reported in the last decade. The intuitive results achieved by pixel-level fusion can benefit the end users through direct observation. Among these pixel-level methods, transform domain- based approaches account for a dominant solution due to the inspiration of the human visual system.[28] The general steps for the transform domain-based multimodal image fusion include transforming the input images to a specific transform domain, performing fusion operation by combining coefficients, and generating the fused image by applying the inverse transform. Various transform methods have been proposed, including stationary wavelet transform,[29] discrete wavelet transform,[30] nonsubsampled contourlet transform,[31] self-fractional Fourier functions,[32] dual-tree complex wavelet transform with sparse representation (DTCWT-SR),[3] convolutional sparse representation (CSR),[4] etc. In addition, fusion operations were also implemented with hand-crafted fusion strategies, such as guided filtering-based weighted average,[33] choose-max,[34] etc. A comprehensive review of the state-of-the-art methods is available in Ref. 25.

It is proved that the fused image with enhanced contents is suitable for human visual perception and low-level image processing. However, the pixel-level fusion has limited benefit for the real-time machine processing and analysis, such as target detection. These image fusion methods are computationally intensive, and the hand-crafted fusion strategies are not able to capture all of the important features from each modality. In contrast to these hand-engineered fusion methodologies, our DIF fuses the complementary information from the multimodal images through a powerful machine learning model, i.e., CNNs. Moreover, the fusion module is integrated into the target detection framework and multimodal learning is trained jointly. In this way, the proposed fusion strategy is implemented through a learning method rather than a manual design. Furthermore, the computation becomes more efficient due to the shared computational resources with the target detection application.

2.3 *Deep Convolutional Neural Networks*

CNNs[35] have brought a series of breakthroughs for many generic computer vision challenges recently, such as image recognition,[36] object detection,[22] and semantic segmentation.[37] CNN is a trainable feedforward neural network mainly comprising convolution layers, pooling layers, and normalization layers. By training on a large-scale dataset, the CNN can learn a hierarchical representation of an object or scene. Recent work reported in Refs. 38 to 40 has shown that a deeper CNN can help gain better performance on computer vision tasks. Driven by this insight, He et al. proposed a ResNet[39] network, which comprised hundreds of convolution layers and broke many records in numerous tasks. Meanwhile, there are a few deep CNNs-based methods for multimodal image fusion. Recently, a CNN-based fusion method for multi-focused images was reported in Ref. 41. The authors from Refs. 42 and 43 made an effort to solve the remote sensing fusion problem with the CNN models. In this study, we leverage the power of deep CNN to fuse visible, IR, and motion images (MIs) and further improve the performance for military target detection.

3 Deep Image Fusion Methodology

3.1 *Overall Framework Description*

The overall illustration of the DIF is shown in Fig. 2. It has four processing steps to obtain the output from the input. The first step is the multimodal image preprocessing, where three different types of images are processed and combined into an RGB-channel image before being fed into the networks. In the second step, the fusion operation and deep feature extraction from the multimodal images through the RGB channels are performed. In the last step, the possible target regions are identified from the deep features derived by the RPN. In the final step, each possible target region is classified and the accurate target bounding box is drafted.

3.2 *Multimodal Image Preprocessing*

In this study, three different types of images are considered in the DIF framework: (1) midwave infrared image (MWIR), (2) gray-scale VI, (3) and MI generated from two consecutive visible frames/images.

3.2.1 *Midwave infrared image*

The MWIR image belongs to the category of the passive IR image in which no external light source is required in comparison with an active IR image. And the electromagnetic spectrum of MWIR image is from 3 to 5 μm. Thus, the MWIR imaging can capture temperature variations over the target and background over a relatively long distance and produce thermograms in the form of a two-dimensional (2-D) image. The value in each coordinate of thermograms represents the relative temperature. To process with the DIF deep feature extraction module, the thermograms need to be transformed into the general gray-scale images by applying the following linear normalization:[44]

$$I = \frac{[T(x,y) - \mathrm{Min}(T)] \times (v_{\max} - v_{\min})}{\mathrm{Max}(T) - \mathrm{Min}(T)} + v_{\min}, \tag{1}$$

where T and I are the 2-D thermogram and gray-scale thermal image, respectively. (x, y) indicates the 2-D coordinates in the image array. $\mathrm{Max}(\cdot)$ and $\mathrm{Min}(\cdot)$ refer to the functions used to obtain the maximum and minimum value among the data. The intensity range of the gray-scale thermal image is $(v_{\min}, v_{\max})$.

3.2.2 *Visible image*

The VI image in this study is of the electromagnetic spectrum range from 380 to 750 nm. This spectral range enables VI to capture sufficient edge and texture information from the scene. However, the disadvantage is that VI is extremely sensitive to the luminance variation. In the experiments, we assume that the VI is aligned with MWIR already and we do not need to perform any registration operations.

3.2.3 *Motion image*

It is well-known that a moving object can generate a motion trajectory. Hence, motion estimation is a straightforward way to obtain the location information of moving targets, even though the associated noises will sometimes be present. DIF leverages an MI modality in the fusion process. Taking the computational complexity into account, an efficient motion estimation method is used to obtain the MI. The method is formulated as follows:

$$M_t = |V_t(x,y) - V_{t-\delta}(x,y)|, \tag{2}$$

where M and V represent the MI and the original VI, respectively. t indicates the t'th frame in a continuous image sequence. δ means the frame interval between two consecutive key frames,

which is affected by the frame sample rate. In our experiments, we adjust the sample rate to control the δ, and (x, y) indicates the 2-D coordinate in the image array.

As illustrated in the multimodal images preprocessing step in Fig. 2, the three obtained images are combined into the RGB-channel of one image. Note that the value of each image has not been modified at this step. The different combination orders of the multimodal images were analyzed in the experiments, and the best combination placed VI in the blue (B) channel, MI in the green (G) channel, and MWIR in the red (R) channel.

3.3 *Deep Feature Extraction from Multimodal Images*

As described in Ref. 45, an image fusion algorithm is used to solve two key problems: (1) effectively extracting the image features from the input source images and (2) combining the features from multiple sources into the fused image. For example, in traditional image fusion, both multiscale transform-based methods[3,29,30,34] and sparse representation-based methods[4] are developed to solve the first problem, feature representation. The fusion strategies, e.g., weighted average[33] and choose-max,[34] are applied to address the second problem, feature fusion. These principles help better understand the proposed deep feature extraction module.

As can be seen in Fig. 3, a set of learnable kernels are convolved on the RGB-channel image, and they generate a set of feature maps. Selecting one kernel as the example, the convolution procedure[35] can be formulated as follows:

$$y_{(i,j)}^{l+1} = \sum_{a=-\frac{m}{2}}^{\frac{m}{2}} \sum_{b=-\frac{m}{2}}^{\frac{m}{2}} \sum_{c=0}^{d-1} w_{(a+\frac{m}{2},b+\frac{m}{2},c)} x_{(i+a,j+b,c)}^{l} + \beta, \tag{3}$$

where l represents the l'th feature layer within the deep CNN. The y is the generated 2-D feature map of the $l+1$'th convolutional layer while x is the original three-dimensional (3-D) feature bank (i.e., the RGB-channel image for the first convolutional layer). (i, j) indicates the 2-D coordinate in the feature map. w is the convolutional kernel/weight with a width and height of m and depth of d. Note that the depth of kernel d should be equal to the channel size of the original feature bank and β is the bias value.

The first 3-D convolution operation in the deep feature extraction procedure is a kind of weighted fusion strategy. Nevertheless, in contrast to the traditional weighted fusion rule, deep feature extraction can learn how to assign the weights to each image modality and extract the important feature from both within-modality and cross-modality.

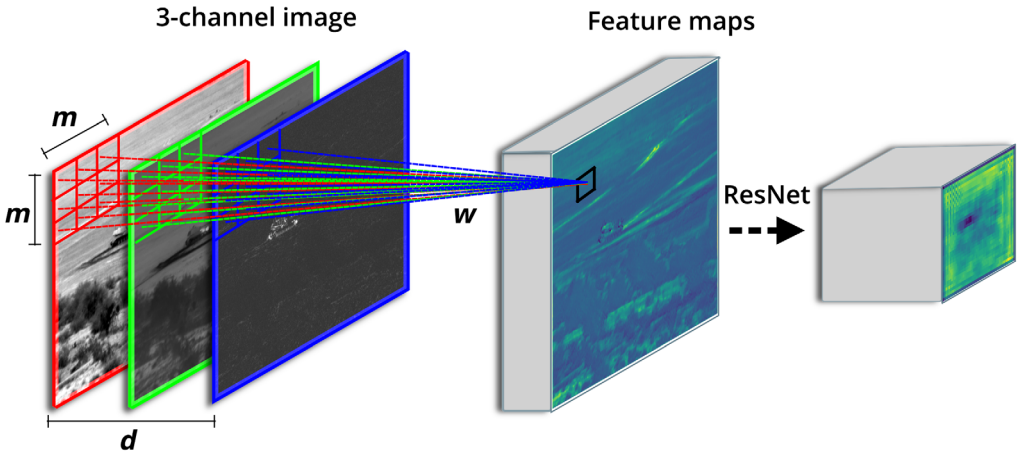

Fig. 3 The illustration of proposed deep feature extraction module. The *m* represents the convolution kernel size, and the *w* indicates the learnable weights of a convolution kernel. The *d* is the number of channels for an image or feature map.

In general, a deep CNN stacks a combination of convolutional layers, activation layers, normalization layers, and pooling layers and repeats this pattern until the spatial scale of feature maps is a small size. With the increase of convolutional layers, the spatial scale of feature map decreases and the produced feature maps become more abstract and well-represented. As a result, the network can learn a hierarchical representation of the multimodal images.

Recent work[38,40] demonstrated that the deeper CNN achieved a better result on representation learning tasks. However, directly increasing the convolutional layers causes a degradation in performance. To address this issue, He et al.[39] proposed a residual block module that allows information to be passed directly through, making the backpropagated error signals less prone to exploding or vanishing. This solution makes it possible to train networks with hundreds of layers. They also carried out a deep CNN model, called ResNet 101, which consists of 101 convolutional layers. In this study, we adopted this state-of-the-art ResNet 101 to fuse multimodal images and extract deep feature maps for the RPN and classification and regression network. We will not duplicate here the materials from Ref. 39 due to the limited number of pages. Readers are referred to Ref. 39 for more details. This work leverages transfer learning for faster training by pretraining the ResNet 101 on a larger-scale image dataset ImageNet.[46] We truncated the pretrained ResNet 101 at the last layer of the "conv4" block and only used the former fully convolutional network for our task. The dilated convolution[47] was also performed to increase the receptive field as in Ref. 48.

3.4 *Target Region Proposal*

The objective of target region proposal is to generate a set of class-independent locations that are likely to contain targets. We adopted the selective search algorithm[13] to accomplish this task in our previous work.[7] As the selective search with a complex implementation can only run on a CPU, it is not efficient for real-time applications. Recently, Ren et al.[14] introduced an RPN that is a fully convolutional network to accelerate the region proposal procedure. The RPN will output a set of rectangular target proposals with corresponding objectness scores and share the convolutional computation with the other networks. In other words, RPN is a small network module that performs region proposal on the last layer of the main deep CNN. The core idea behind the RPN is the anchors. Specifically, anchors are a set of reference boxes with different scales and ratios on a regular grid in the image. The generated region proposals are the offsets to the anchors, and thus the number of region proposals is fixed.

The configuration of RPN network is shown in Fig. 4. To be specific, a 3×3 convolution layer with Relu[49] activation function slides on the feature maps generated by ResNet 101, followed by two sibling 1×1 convolution layers, e.g., one is for outputting region proposals and the other is for outputting the corresponding objectness scores. Readers are referred to Ref. 14 for details on the loss function and implementation.

3.5 *Classification and Localization*

As shown in Fig. 5, the regionwise classification and regression network is applied to each region proposal and will generate classification scores as well as four offset values with respect to

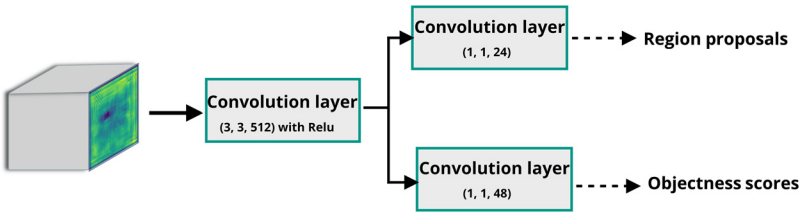

Fig. 4 The network configuration of RPN module. For the setting of the convolution layer, "(size, size, number)" denotes the width, height, and number of convolution kernels. The "with Relu" means that the convolution layer is followed by an activation function of rectified linear unit (Relu).[49]

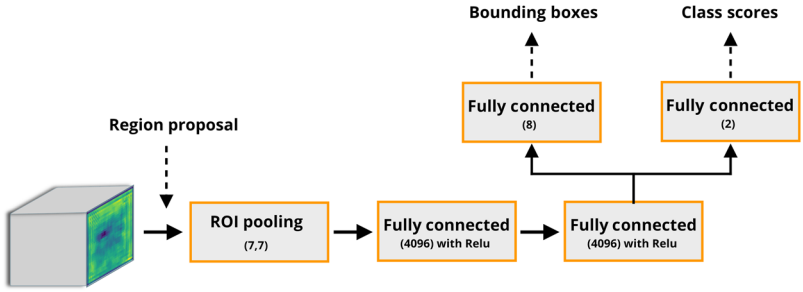

Fig. 5 The network configuration of the regionwise classification and regression network. For the configuration of ROI pooling, the "(size,size)" denotes the width and height of the pooling kernel. For the configuration of the fully connected layer, the "(number)" represents the number of output neurons. The fully connected layer with Relu means that a Relu activation function is followed by the layer.

the bounding box of the region proposals. Hence, this network has two functional heads, e.g., classification and regression. The first one is to classify the region proposals and output a discrete probability value over two categories (target and background) using the Softmax[50] function. The second one is to regress the bounding box offsets of the region proposals and output a tuple of (t_x, t_y, t_w, t_h), where the elements indicate the shift value relative to the central coordinate, height, and width of the original region proposal.

To train the classification head, the cross entropy is used as the loss function:

$$L_{\text{cls}}(p, u) = u \log(p) + (1 - u) \log(1 - p), \tag{4}$$

where p and u represent the ground-truth label of the target/background and the predicted probability, respectively. Meanwhile, the smooth L_1 loss function[8] is adopted as the loss function for the regression head

$$L_{\text{bbox}}(t^u, v) = \sum_{i \in \{x,y,w,h\}} \text{smooth}_{L_1}(t_i^u - v_i), \tag{5}$$

in which $\text{smooth}_{L_1}(x)$ can be expressed as

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 0 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \tag{6}$$

where t^u is the bounding box offsets of the u class. And v is the true offsets.

At the training stage, both of the two loss functions will be put together as in[8]

$$L(p, u, t^u, v) = L_{\text{cls}}(p, u) + \lambda[u = 1] L_{\text{bbox}}(t^u, v), \tag{7}$$

where $u = 1$ means only when the class is a target, the bounding box regression can be trained, and λ is used to control balance between classification and regression. λ is set to 1 in all of the experiments.

4 Experimental Results

4.1 *Dataset*

The large-scale military image datasets are not accessible to the public research community. Recently, several unclassified military datasets were available for research use including SENSIAC[51] and DARPA VIVID.[52] We evaluated our proposed approach on the ATR dataset from the Military Sensing Information Analysis Center (SENSIAC). This dataset contains 207 GB of MWIR imagery (video) and 106 GB of visible imagery (video) along with ground

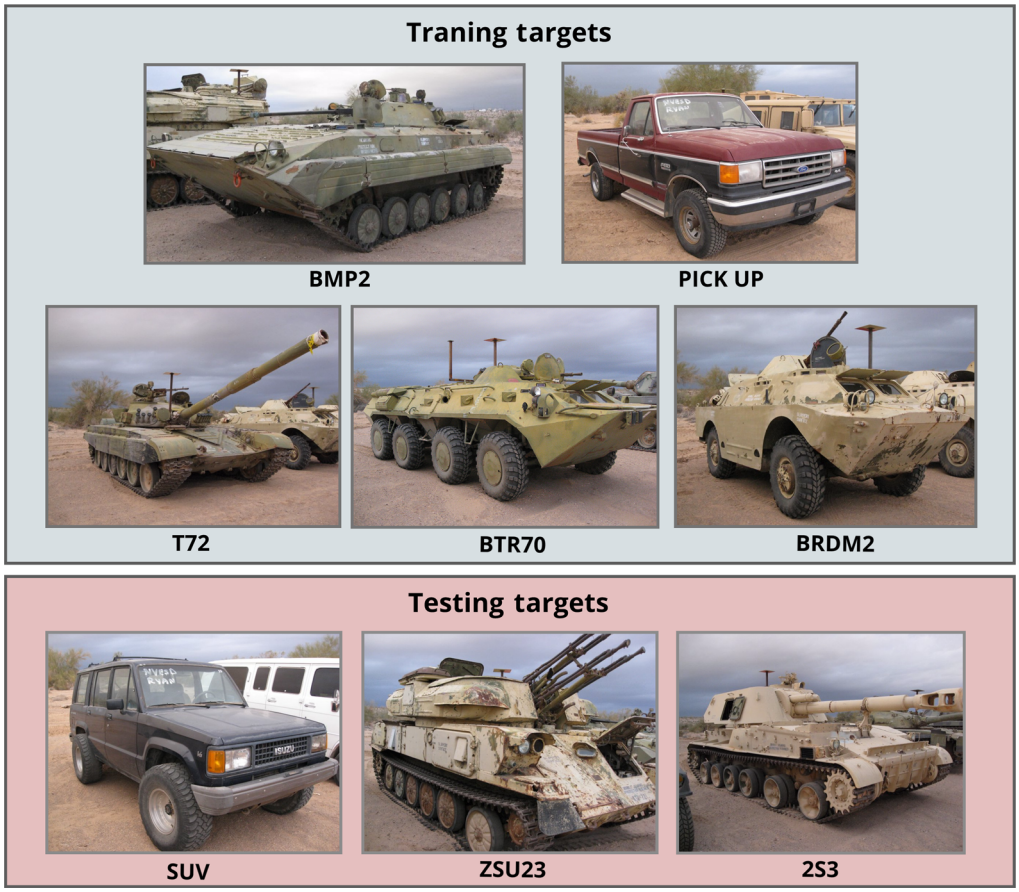

Fig. 6 Appearance of targets in training data and testing data.

truth data. All imagery was taken using commercial cameras operating in the MWIR and VI bands. Various types of objects are included in this dataset, for instance, soldiers, military vehicles, and civilian vehicles. Moreover, the dataset was collected during both the daytime and nighttime with multiple observation distances (ODs) from 500 to 5000 m.

In the experiments, we only considered the vehicle (ignoring its type) as the target. As shown in Fig. 6, we categorized five types of vehicles into training targets and three new types of vehicles into testing targets to examine the generalization of the trained models. In the first comparative experiment, we selected three different ODs (1000, 1500, and 2000 m) as in Ref. 7 and sampled the key frame at 6 Hz (every five frames). So we had 4573 training images and 2812 testing images. For the subsequent experiments, we selected nine different ODs (long distances) from 1000, 1500, to 5000 m. To further reduce the overall data size in the experiments, the sample rate was reduced to 3 Hz. Eventually, there were 7688 training images and 3542 testing images in total.

4.2 *Experimental Setup*

The proposed DIF system was implemented using Tensorflow deep learning toolbox.[53] For the training, we used a machine with an NVIDIA GeForce GTX 1080 GPU, an Intel Core i7 CPU, and 32 GB memory. For the hyperparameters, we trained each of the networks for 60,000 iterations with initial learning rate 0.0003 and 0.00003 for the last 40,000 iterations, with a batch size 1, momentum 0.9, and weight decay 0.0005. In addition, all of the newly added layers were initialized from a Gaussian distribution with zero mean and 0.001 variance.

We selected the *de-facto* standard average precision (AP) as the evaluation metric, which is calculated as the ratio between the area under the precision–recall curve and the entire area (which is 1).

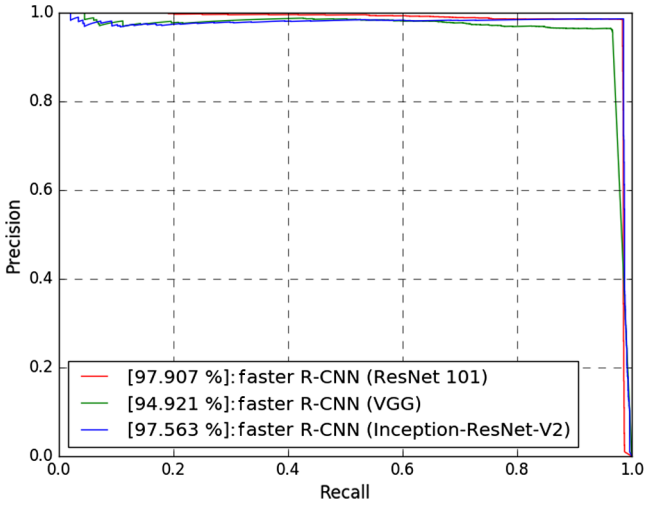

Fig. 7 The accuracy comparison of different CNNs and target detection architecture.

4.3 *System and Performance Optimization*

4.3.1 *CNN architectures*

The proposed DIF is built based on the faster R-CNN.[14] In this section, we investigate the different popular CNN architectures for the faster R-CNN; they are VGG,[38] inception-ResNet-V2,[54] and ResNet 101.[39] Note that the original faster R-CNN was coupled with the VGG. As can be seen in Fig. 7, the faster R-CNN coupled with the ResNet 101 was much better than that coupled with the VGG, with around 3% boosted accuracy. Thus, we adopted the ResNet 101 as the base CNN architecture in the DIF and set the faser R-CNN with ResNet 101 as the baseline in the experiments.

4.3.2 *Modal orders*

In the DIF, the three different modalities (VI, MWIR, and VI) are combined into the RGB-channel of one image before being fed into the neural network. In this section, we examine the different orders of the modalities in the RGB-channel image. In Fig. 8, the MI-MWIR-VI means that MI, MWIR, and VI were put into the red, green, and blue channels of the composite

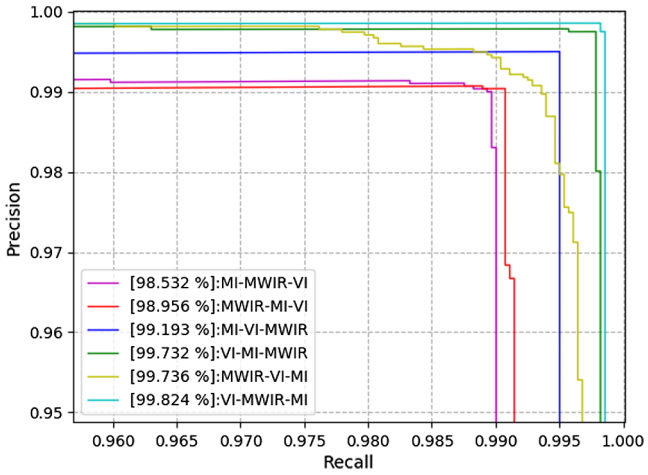

Fig. 8 The accuracy comparison of different modal orders.

image, and the other permutation-and-combinations followed this format. As can be seen in Fig. 8, all of the possible permutation-and-combinations were enumerated. The VI-MWIR-MI combination achieved the best performance with (99.824), which is almost the ceiling performance.

4.4 *Comparison with the State-of-the-Arts*

To our best knowledge, there is no publicly available system that focuses on image fusion-based target detection. Thus, it is a challenge to compare with the same-level approaches on the SENSIAC dataset. As mentioned above, the DIF is built based on the faster R-CNN, but the faster R-CNN is only able to process the VIs. So we set the faster R-CNN as our baseline method to validate the benefits from the DIF. We also compared the proposed DIF system with the conventional image fusion methods (DTCWT-SR[3] and CSR[4]). These methods were designed for visualization not target detection. For a fair comparison, we applied these methods to fuse VI and MWIR images under the default configurations and then fed the fused images to the faster R-CNN (ResNet 101) for target detection. Our previous work[7] was also compared in this experiment. We reported two versions of the DIF, VI-MWIR, and VI-MWIR-MI. The VI-MWIR represents the fusion of VI and MWIR, and the VI-MWIR-MI represents the fusion of VI, MWIR, and MI.

The accuracy comparison results are presented in Fig. 9. Compared with the baseline method faster R-CNN (ResNet 101), the DIF has a 1.917% AP improvement. In comparison with the state-of-the-art hand-engineered image fusion methods, DTCWT-SR and CSR, DIF (VI-MWIR) reached 0.590% and 0.642% improvements; this means that our DIF is able to learn a better strategy on assigning the weights to each image modality and choosing the important cross-modality information compared with those hand-engineered strategies. When we added the motion modality into the DIF, the DIF (VI-MWIR-MI) gained a 0.355% improvement compared with the DIF (VI-MWIR), and the proposed DIF method also outperformed the previous work.

Another important consideration is run-time efficiency. We reported the efficiency comparison results in Table 1. The DIF (VI-MWIR) took only 0.238 s to process an image, which is around 6× faster than our previous work and over on order of magnitude faster than other conventional image fusion-based methods. The reason is that the neural networks module can be easily optimized on a GPU device, where the main computational cost in a multimodal image fusion-based detection system is with the image fusion module and the region proposal module. For instance, the CSR + faster R-CNN (ResNet 101) method cost is 6.179 s on the image fusion module, and our previous work required 1.272 s on the region proposal module. But the proposed DIF method combines those three modules into an end-to-end neural network, which enables them to share the computational resources and be optimized on a GPU device. The DIF (VI-MWIR-MI) increased 0.355% in accuracy but only dropped 0.018 s running time compared with the DIF (VI-MWIR), which is an acceptable trade-off.

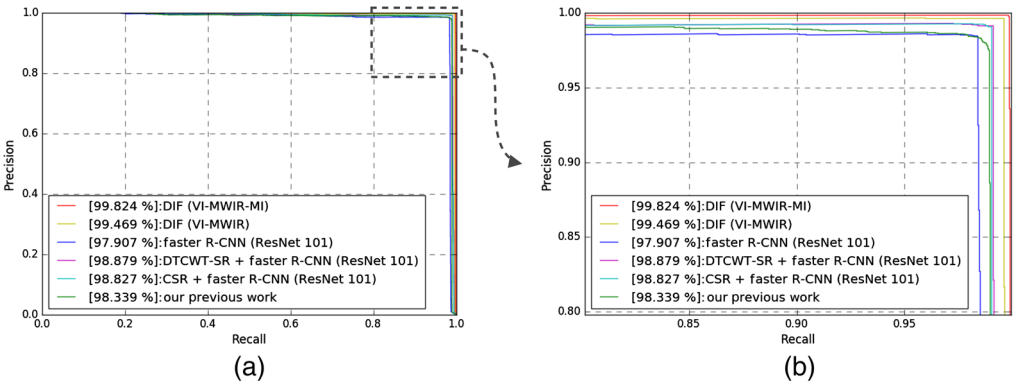

Fig. 9 Comparison of the state-of-the-art methods. (a) The overall precision–recall curves of different methods. (b) The local enlarged image from (a).

Table 1 Performance comparison of time cost of different multimodal image fusion-based methods.

Methods	Running time (second/image)	AP (%)
Our previous work	1.507	98.339
Faster RCNN (ResNet 101)	0.242	97.907
CSR + faster R-CNN (ResNet 101)	6.413	98.827
DTCWT-SR + faster R-CNN (ResNet 101)	2.758	98.879
DIF (VI-MWIR)	**0.238**	99.469
DIF (VI-MWIR-MI)	0.256	**99.824**

Note: The best running time and average precision are both highlighted in bold font.

4.5 *Analysis and Discussion*

4.5.1 *Target scales*

As described in Sec. 1, there are many factors that degrade the performance of target detection. One critical factor is the target scale, i.e., the OD between the imaging system and the target, especially in a complex scenario. In this section, experiments were conducted for comprehensive analysis of the multiscale situation with the SENSIAC dataset.

Note that there is an inverse relationship between the target scale and the OD from the imaging system to the target. This means that a longer OD leads to a smaller target scale. For the sake of simplicity, we utilize the "observation distance" term to represent the relative target scale in our experiments.

We selected a set of data across a long OD (from 1000 to 5000 m) and trained the detector with all of the selected data. For evaluation, we assessed the detection performance for targets at different ODs. The OD range of [1000, 2000] m is classified as the large target scale while the [2500, 5000] m range is the small target scale. To verify the effectiveness of the DIF method, we implemented five detectors of incremental image modality, from single image modality (MWIR, VI, and MI) to multimodal image fusion (MWIR-VI and MWIR-VI-MI).

Figure 10 shows the AP results against ODs for different modalities in the small OD. In general, VI-MWIR-MI and VI-MWIR performed better than other modal combinations. The MI modality had an unsatisfying overall performance, and it also degraded the performance of VI-MI and MWIR-MI in the most distances.

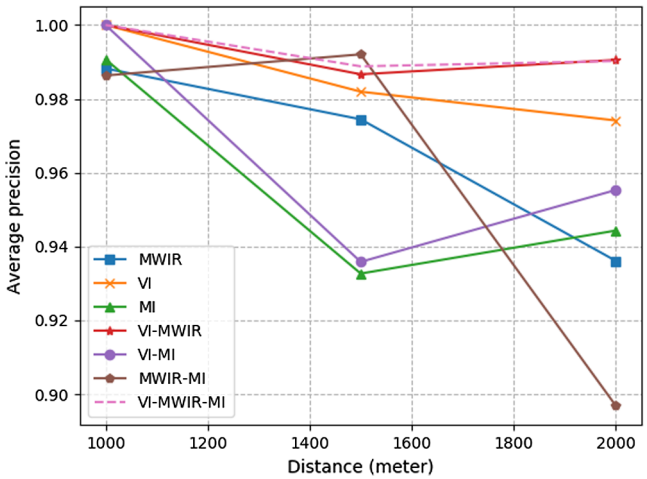

Fig. 10 The AP comparison against OD in large target scales.

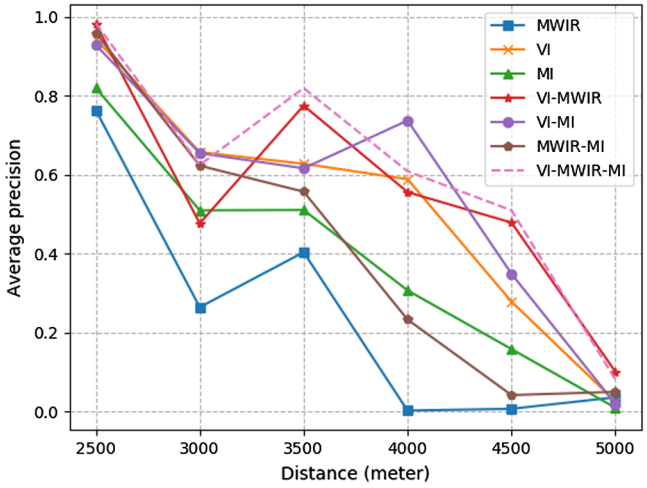

Fig. 11 The AP comparison against OD in small target scales.

We further compared different methods at the long ODs. Figure 11 shows that the performance of all types of image modalities decreased significantly with the increase of OD. Similar to the results on the close OD test, the VI-MWIR-MI and VI-MWIR performed better than other modal combinations, especially for the extremely small target (in 4500 and 5000 m). It is worth mentioning that, even with the target that is 3500 m away, the DIF of MWIR, VI, and MI can still achieve 81.9% AP for the detection. However, we found that the MWIR modality performed extremely poorly at 3000 m and almost all of the modalities were worse at 3000 m than that at 3500 m. Hence, this raises a question about the existence of other critical factors affecting the performance of the detection. We will describe what we discovered in the next section.

4.5.2 *Environmental complexity*

In a complex scene, a cluster of trees and rocks are the ideal natural covers to obscure targeted objects, which introduces a challenge for target detection. We define this critical impact factor as "environmental complexity." To measure this factor, we first define the signal and noise for the scenario of target detection in Fig. 12. The red dash bounding box is the target area, which is the system to detect, so we treat it as the signal. Then, we increase the target bounding box by $\sqrt{2}$ to the green bounding box. Thus, the area between the red and the green bounding box refers to the local environment area. From the observation, we found that the noise factors appearing within the local environment degraded the performance of target detection, so we set the local environment area as the noise. To quantify the environmental complexity, we calculated the signal-to-noise ratio (SNR)[55] as follows:

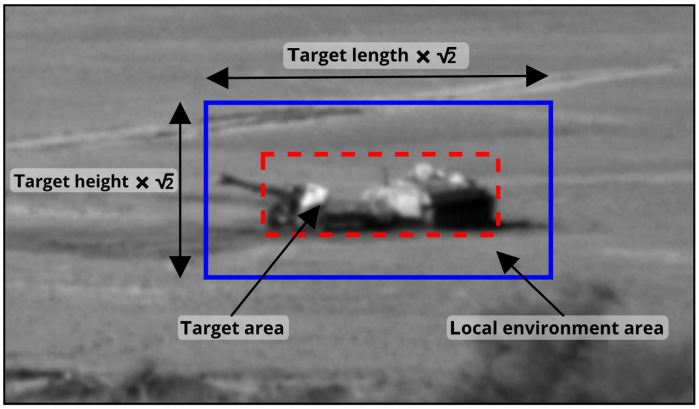

Fig. 12 Illustration of target area vs. local environment area.

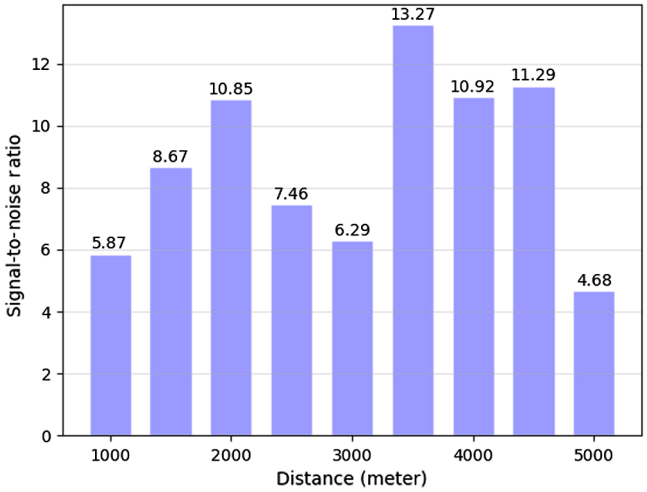

Fig. 13 The distribution of SNR value of MWIR imageries against distances.

$$\mathrm{SNR} = \frac{\mu_{\mathrm{signal}}}{\sigma_{\mathrm{noise}}}, \tag{8}$$

where μ_{signal} is the mean value of the signal and σ_{noise} is the standard deviation of the noise. When there are noises in the local environment, e.g., cluster of trees or rocks, the σ_{noise} will increase and reduce the SNR value. Therefore, a higher SNR score indicates lower environmental complexity.

The SNR distribution of MWIR imageries against ODs is shown in Fig. 13. The SNR score at 3000 m is 1.17 lower than that at 2500 m and almost half of the SNR value at 3500 m. In other words, the natural environment at 3000 m is much more complex than its neighbors. This explains why the detection performance by the MWIR modality at 3000 m is worse than the other methods in Fig. 11.

4.5.3 *Statistical analysis*

As discussed in the above section, there are two factors, e.g., target scale (OD) and environment complexity, that are critical to the target detection performance. In this section, we seek to employ a statistical method to verify if the proposed deep fused system will mitigate the impact of the target scale and environmental complexity. To accomplish this, the SNR, OD, and corresponding AP for different image modalities were calculated, and the results are summarized in Table 2.

Multiple linear regression analysis is used to evaluate the association between two or more independent variables and a dependent/response variable. In this study, we set the OD and environmental complexity (SNR) as two independent variables and the performance of detectors (AP) as the dependent variable. Hence, we formulate the linear regression model as follows:

$$\mathrm{AP} = b_0 + b_1 \times \mathrm{OD} + b_2 \times \mathrm{SNR}, \tag{9}$$

where b_1 to b_2 are the estimated coefficients and b_0 is a constant term. The multiple linear regression presents the equation that minimizes the distance between the fitted line and all of the data points. If the two factors, e.g., target scale and environmental complexity, have a great influence on the detection, the estimated multiple linear regression model will fit the data well. In other words, the lower goodness-of-fit for a multiple linear regression means that the detection system has less dependence on the OD (target scale) and environmental complexity. So a low goodness-of-fit is the expectation. To measure the goodness-of-fit of the model, three well-known statistical metrics (R^2, adjusted R^2, and p-value) were adopted. R^2, also called the coefficient of determination, is a statistical measure of how close the data are to the fitted regression line. The value of R^2 is in the range [0, 1]. A higher value means that the multiple linear regression has a better goodness-of-fit but the target detection modal has more dependence on

Table 2 Distance, SNR, and corresponding AP for different image models.

MWIR			VI			MI			MWIR-VI			VI-MI			MWIR-MI			VI-MWIR-MI		
OD	SNR	AP	OD	SNR	AP	OD	SNR	AP	OD	SNR	AP	OD	SNR	AP	OD	SNR	AP	OD	SNR	AP
1000	5.865	0.988	1000	3.413	1.0	1000	2.726	0.990	1000	4.231	1.0	1000	1.191	1.0	1000	1.449	0.986	1000	1.285	1.0
1500	8.666	0.974	1500	4.799	0.981	1500	2.813	0.932	1500	6.088	0.986	1500	1.329	0.935	1500	1.448	0.992	1500	1.309	0.989
2000	10.848	0.936	2000	4.759	0.974	2000	2.727	0.944	2000	6.788	0.990	2000	1.161	0.955	2000	1.507	0.896	2000	1.249	0.990
2500	7.460	0.763	2500	4.568	0.944	2500	2.636	0.820	2500	5.532	0.981	2500	1.192	0.929	2500	1.445	0.959	2500	1.294	0.984
3000	6.294	0.264	3000	6.831	0.657	3000	2.632	0.509	3000	6.652	0.476	3000	1.484	0.654	3000	1.333	0.623	3000	1.248	0.625
3500	13.265	0.403	3500	7.069	0.627	3500	2.536	0.510	3500	9.134	0.775	3500	1.261	0.616	3500	1.332	0.557	3500	1.288	0.819
4000	10.919	0.002	4000	5.995	0.588	4000	2.357	0.306	4000	7.636	0.555	4000	1.363	0.737	4000	1.463	0.232	4000	1.353	0.608
4500	11.294	0.006	4500	7.449	0.278	4500	2.524	0.157	4500	8.731	0.478	4500	1.658	0.348	4500	1.437	0.041	4500	1.499	0.509
5000	4.677	0.035	5000	6.925	0.025	5000	2.138	0.009	5000	6.176	0.100	5000	1.792	0.018	5000	1.536	0.049	5000	1.586	0.081

Table 3 Results of multiple linear regression for the data in Table 2.

	MWIR	VI	MI	VI-MWIR	VI-MI	MWIR-MI	VI-MWIR-MI
R^2	0.8927	0.8823	0.9573	0.8595	0.9314	0.9126	**0.8558**
Adjusted R^2	0.8570	0.8431	0.9431	0.8127	0.9085	0.8835	**0.8077**
p-Value	0.0012	0.0016	7.769e-05	0.0027	0.0003	0.0007	**0.0030**

Note: The best results are highlighted in bold font.

the noise factors. The adjusted R^2 is a modified version of R^2, which has one more term that penalizes a model for each additional explanatory variable. Consequently, any variable without a strong correlation will make the adjusted R^2 decrease. The p-value is used to test the null hypothesis that the independent variables (i.e., target scale and environmental complexity) have no effect on the response variable (i.e., AP). So in this case, a higher p-value indicates that it is more possible to accept the null hypothesis. In other words, the regression model with a higher p-value means that the detection system has less dependence on the OD (target scale) and environmental complexity.

A set of multiple linear regression models were estimated for detectors of different image modalities. The results are given in Table 3. The overall evaluation showed that all of the multiple linear regression models of the detectors have a high goodness-of-fit, which means that the target scale and environmental complexity have a strong effect on the performance of the detectors. By contrast, the proposed DIF system (MWIR-VI-MI) had the lowest values for both R^2 and adjusted R^2 and the highest p-value compared with the single modal-based and double modal-based methods. This means that the DIF method is able to alleviate the impact of the target scale and environmental complexity in comparison with the single image modalities.

4.6 *Summary of Analysis*

The experiments for algorithm comparison demonstrate both the effectiveness and efficiency of the proposed framework for deep multimodal image fusion. In the analysis of OD, the DIF performs better than other unimodal methods, especially in a long OD. However, when any individual imaging modality involved in the fusion framework has a degraded performance, it will also introduce degradation to the overall deep multimodal detection.

Two factors, i.e., OD (target scale) and environment complexity, were investigated for their impacts on detection performance. As the target becomes smaller in a longer distance, the detection performance will get worse generally. Another observation is that lower environmental complexity will allow a better result in the detection. When taking both factors into account, the statistical analysis showed the evidence that the proposed DIF can significantly mitigate the two impacts.

5 Conclusions

In this paper, we proposed a CNN-based DIF framework for target detection in complex battlefields. The capability of detecting small targets in a complex environment will enhance the real-time situation awareness in a battlefield. The overall framework configured in an end-to-end network is composed of multimodal image preprocessing, deep feature extraction/fusion, region proposal, classification, and regression modules. The extensive experiments on the SENSIAC dataset demonstrated that the proposed method achieved 99.82% accuracy with great computational efficiency for real-time applications. Moreover, the proposed fusion method can deal with varied noises from a complex background. Thus, the DIF framework has great potential to function in a real world application. The SENSIAC is so far the most comprehensive dataset with multimodal images for target detection, but the sample images of different vehicles are still limited for target classification research. For future work, we plan to apply our DIF method to more available datasets and enable target classification through DIF as well.

Acknowledgments

The authors wish to thank the editors and the reviewers for constructive criticism that resulted in significant improvements to the paper.

References

1. A. C. Muller and S. Narayanan, "Cognitively-engineered multisensor image fusion for military applications," *Inf. Fusion* **10**, 137–149 (2009).
2. Z. Wenda, L. Huimin, and W. Dong, "Multisensor image fusion and enhancement in spectral total variation domain," *IEEE Trans. Multimedia* **20**(4), 866–879 (2017).
3. Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion* **24**, 147–164 (2015).
4. Y. Liu et al., "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.* **23**, 1882–1886 (2016).
5. H. Hai-Miao et al., "An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels," *IEEE Trans. Multimedia* **19**, 2706–2719 (2017).
6. G. Jie, M. Zhenjiang, and Z. Xiao-Ping, "Efficient heuristic methods for multimodal fusion and concept fusion in video concept detection," *IEEE Trans. Multimedia* **17**, 498–511 (2015).
7. S. Liu and Z. Liu, "Multi-channel CNN-based object detection for enhanced situation awareness," arXiv:1712.00075 (2017).
8. R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, Boston, Massachusetts, pp. 1440–1448 (2015).
9. J. Gong et al., "Joint view-identity manifold for infrared target tracking and recognition," *Comput. Vision Image Understanding* **118**, 211–224 (2014).
10. H. Chen et al., "Task-driven progressive part localization for fine-grained object recognition," *IEEE Trans. Multimedia* **18**, 2372–2383 (2016).
11. L. Xinchen et al., "Provid: progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Trans. Multimedia* **20**, 645–658 (2018).
12. L. Yu et al., "Joint infrared target recognition and segmentation using a shape manifold-aware level set," *Sensors* **15**, 10118–10145 (2015).
13. J. R. R. Uijlings et al., "Selective search for object recognition," *Int. J. Comput. Vision* **104**, 154–171 (2013).
14. S. Ren et al., "Faster R-CNN: towards real-time object detection with region proposal networks," in *Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, pp. 91–99 (2015).
15. E. Blasch, C. Yang, and I. Kadar, "Summary of tracking and identification methods," *Proc. SPIE* **9091**, 909104 (2014).
16. E. Gundogdu et al., "Comparison of infrared and visible imagery for object tracking: toward trackers with superior IR performance," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. Workshops*, IEEE, Boston, Massachusetts, pp. 1–9 (2015).
17. H. S. Demir and A. E. Cetin, "Co-difference based object tracking algorithm for infrared videos," in *IEEE Int. Conf. Image Process.*, IEEE, Phoenix, Arizona, pp. 434–438 (2016).
18. J. Gong et al., "Infrared target tracking, recognition and segmentation using shape-aware level set," in *IEEE Int. Conf. Image Process.*, IEEE, Melbourne, Victoria, pp. 3283–3287 (2013).
19. B. Millikan et al., "Fast detection of compressively sensed IR targets using stochastically trained least squares and compressed quadratic correlation filters," *IEEE Trans. Aerosp. Electron. Syst.*. **53**, 2449–2461 (2017).
20. P. F. Felzenszwalb et al., "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1627–1645 (2010).
21. P. Sermanet et al., "OverFeat: integrated recognition, localization and detection using convolutional networks," arXiv:1312.6229 (2013).
22. R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, Columbus, Ohio, pp. 580–587 (2014).

23. K. He et al., "Spatial pyramid pooling in deep convolutional networks for visual recognition," *Lect. Notes Comput. Sci.* **8691**, 346–361 (2014).
24. W. Liu et al., "SSD: single shot multibox detector," *Lect. Notes Comput. Sci.* **9905**, 21–37 (2016).
25. Z. Liu et al., "Fusing synergistic information from multi-sensor images: an overview from implementation to performance assessment," *Inf. Fusion* **42**, 127–145 (2018).
26. R. S. Blum and Z. Liu, *Multi-Sensor Image Fusion and Its Applications*, Signal Processing and Communications, 1st ed., Taylor and Francis, Hoboken, New Jersey (2005).
27. Y. Zheng and E. Blasch, "Multispectral image fusion for vehicle identification and threat analysis," *Proc. SPIE* **9871**, 98710G (2016).
28. B. A. Olshausen et al., "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature* **381**, 607–609 (1996).
29. M. Beaulieu, S. Foucher, and L. Gagnon, "Multi-spectral image resolution refinement using stationary wavelet transform," in *IEEE Int. Geosci. and Remote Sens. Symp. Proc.*, IEEE, Toulouse, France, Vol. 6, pp. 4032–4034 (2003).
30. A. Loza et al., "Non-Gaussian model-based fusion of noisy images in the wavelet domain," *Comput. Vision Image Understanding* **114**, 54–65 (2010).
31. B. Gaurav, W. Q. M. Jonathan, and L. Zheng, "Directive contrast based multimodal medical image fusion in NSCT domain," *IEEE Trans. Multimedia* **15**, 1014–1024 (2013).
32. K. K. Sharma and M. Sharma, "Image fusion based on image decomposition using self-fractional Fourier functions," *Signal Image Video Process.* **8**, 1335–1344 (2014).
33. S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.* **22**, 2864–2875 (2013).
34. J. Hu and S. Li, "The multiscale directional bilateral filter and its application to multisensor image fusion," *Inf. Fusion* **13**, 196–206 (2012).
35. Y. LeCun et al., "Gradient-based learning applied to document recognition," *Proc. IEEE* **86**, 2278–2324 (1998).
36. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.*, Lake Tahoe, Nevada, pp. 1097–1105 (2012).
37. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, IEEE, Boston, Massachusetts, pp. 3431–3440 (2015).
38. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2014).
39. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, Seattle, Washington, pp. 770–778 (2016).
40. C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, Boston, Massachusetts, pp. 1–9 (2015).
41. Y. Liu et al., "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion* **36**, 191–207 (2017).
42. J. Zhong et al., "Remote sensing image fusion with convolutional neural network," *Sens. Imaging* **17**, 10 (2016).
43. Y. Chen et al., "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.* **54**(10), 6232–6251 (2016).
44. R. C. Gonzales and R. E. Woods, "Digital image processing," (2002).
45. J. Wang et al., "Fusion method for infrared and visible images by using non-negative sparse representation," *Infrared Phys. Technol.* **67**, 477–489 (2014).
46. O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vision* **115**, 211–252 (2015).
47. F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv: 1511.07122 (2015).
48. J. Huang et al., "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 7310–7311 (2017).

49. V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, pp. 807–814 (2010).
50. N. M. Nasrabadi, "Pattern recognition and machine learning," *J. Electron. Imaging* **16**, 049901 (2007).
51. D. Shumaker, "SENSIAC (Military Sensing Information Analysis Center) for PEOs/PMs," Military Sensing Information Analysis Center Atlanta GA, 2008, https://apps.dtic.mil/dtic/tr/fulltext/u2/a535651.pdf (accessed 1 November 2017).
52. R. Collins, X. Zhou, and S. K. Teh, "An open source tracking testbed and evaluation web site," in *IEEE Int. Workshop Perform. Eval. Tracking and Surveillance*, Pittsburgh, Pennsylvania, Vol. 35 (2005).
53. M. A. Abadi et al., "TensorFlow: large-scale machine learning on heterogeneous systems," 2015, tensorflow.org
54. C. Szegedy et al., "Inception-v4, inception-RESNET and the impact of residual connections on learning," in *Proc. Thirty-First AAAI Conf. Artif. Intell.*, Vol. 4, p. 12 (2017).
55. D. J. Schroeder, "Detectors, signal-to-noise, and detection limits," Chapter 17 in *Astronomical Optics*, D. J. Schroeder, Ed., 2nd ed., pp. 425–443, Academic Press, San Diego, California (2000).

Shuo Liu received his MS degree in electrical engineering and computer science from the School of Engineering, University of British Columbia Okanagan Campus, Kelowna, Canada, in 2019. He is currently a data scientist at Two Hat Security Research Ltd., Kelowna, BC, Canada. His research interests include computer vision, machine learning, image fusion, and image translation.

Huan Liu received his PhD in geodetection and information technology from the Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan, China, in 2018. From 2016 to 2017, he was a joint PhD student in electrical engineering and computer science at the School of Engineering, University of British Columbia, Kelowna, Canada. He has been involved in developing intelligent geophysical instruments, especially, the proton magnetometer and the Overhauser magnetometer. He is currently an associate professor in the School of Automation, China University of Geosciences, Wuhan, China. His current research interests include weak magnetic detection, signal processing, data mining, and machine learning. He is a member of IEEE, SPIE, CAA, and a member of the IEEE IMS technical committee on "environmental measurements" (TC-18).

Vijay John received his MS degree in information technology and robotics from Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, in 2005 and his PhD in computer vision from the University of Dundee, Dundee, United Kingdom, in 2007. In 2011, he joined the University of Amsterdam, Amsterdam, The Netherlands, as a postdoctoral researcher. In 2012, he joined Philips Research, Eindhoven, The Netherlands, as a visiting researcher. Since 2014, he has been a postdoctoral research fellow with the Toyota Technological Institute, Nagoya, Japan. His research interests include computer vision, machine learning with applications in autonomous vehicles, and human motion analysis.

Zheng Liu received his doctorate's degree in engineering (earth resources) from Kyoto University, Kyoto, Japan, in 2000 and his PhD in electrical engineering from the University of Ottawa, Canada, in 2007. From 2000 to 2001, he was a research fellow with the Nanyang Technological University, Singapore. He then joined the National Research Council of Canada, Ottawa, Ontario, as a governmental laboratory visiting fellow nominated by NSERC in 2001. From 2002, he was a research officer associated two research institutes of NRC (Aerospace [IAR] and Construction [IRC]). From 2012 to 2015, he worked as a full professor with the Toyota Technological Institute, Nagoya, Japan. He is now with the Faculty of Applied Science at the University of British Columbia—Okanagan as an associate professor. His research interests include predictive maintenance, data/information fusion, computer/machine vision, machine learning, smart sensor and industrial IoT, and nondestructive inspection and evaluation. He is a senior member of IEEE and SPIE.

Erik Blasch is a principal officer at the Air Force Office of Scientific Research in Arlington, Virginia, USA. Previously, he was with the Information Direction in Rome, New York, USA, from 2012 to 2017, and an exchange scientist to Defence Research and Development Canada at Valcartier, Quebec from 2009 to 2012. From 2000 to 2009, he was the information fusion evaluation tech lead for the AFRL Sensors Directorate—Comprehensive Performance Assessment of Sensor Exploitation Center supporting design evaluations in Dayton, Ohio, USA. He is a fellow of SPIE, associate fellow of AIAA, and fellow of IEEE.