Background Subtraction Based on GAN and Domain Adaptation for VHR Optical Remote Sensing Videos

WENTAO YU, (Student Member, IEEE), JING BAI, (Member, IEEE), AND LICHENG JIAO, (Fellow, IEEE)
School of Artificial Intelligence, Xidian University, Xi’an 710071, China
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xi’an 710071, China
Corresponding author: Jing Bai (baijing@mail.xidian.edu.cn)

ABSTRACT The application of deep learning techniques in background subtraction for VHR optical remote sensing videos holds the potential to facilitate multiple intelligent remote sensing processing tasks. However, existing methods on background subtraction for VHR optical remote sensing videos are still facing technical challenges. First, conventional CNN and other networks are limited by performance constraints. Second, existing background subtraction methods are mostly trained by natural videos due to the lack of VHR optical remote sensing video datasets. Third, VHR optical remote sensing videos have large scene sizes. In our article, we design a novel deep learning network via fully utilizing GAN and domain adaptation, which has the ability to measure and minimize the discrepancy between feature distributions of natural videos and VHR optical remote sensing videos so that the background subtraction performance for VHR optical remote sensing videos is improved significantly. Numerous experiments are conducted on the CDnet 2014 dataset and VHR optical remote sensing video dataset. Tremendous experiments demonstrate that our proposed method achieves an average FM of 0.8533, which reveals excellent performance on background subtraction.

INDEX TERMS Background subtraction, generative adversarial networks (GAN), domain adaptation, very high resolution (VHR) optical remote sensing videos.

I. INTRODUCTION

Recently, benefiting from rapidly increasing very high resolution (VHR) optical remote sensing videos, VHR optical remote sensing videos have been widely utilized in various fields. Background subtraction in VHR optical remote sensing videos plays a vital role in military and civil applications, such as military reconnaissance [1], [2], maritime surveillance [3], [4], urban planning & urban construction [5] and traffic monitoring [6], [7]. It is regarded as a crucial component and the key link of intelligent remote sensing processing.

Since considerable attention has been focused on background subtraction in recent years, promising methods have been proposed. Garcia-Garcia et al. [8] presented an exhaustive review on background subtraction. Background subtraction methods based on robust principal component analysis (RPCA) models [9], fuzzy models [10], semantic models [11] and deep learning models [12] are considered promising solutions. Among these methods, deep learning-based methods show great potential in practice owing to their strong generalization capability.

Although deep learning-based background subtraction methods tap into the advantages of deep features and achieve very good performance, there are some technical challenges for VHR optical remote sensing videos. The first technical challenge is that in the application of background subtraction for VHR optical remote sensing videos, features must be extracted more efficiently by stronger generalization capability. However, conventional convolutional neural networks (CNN) and other networks are limited by performance constraints, which do not satisfy our demand.
The acquisition and annotation of VHR optical remote sensing videos are quite expensive and complicated. The second technical challenge is that existing background subtraction methods are almost trained by natural videos. However, natural videos and VHR optical remote sensing videos share similar but different distributions. When these deep learning-based methods are performed on VHR optical remote sensing videos, their performances decrease sharply.

The third technical challenge is that VHR optical remote sensing videos have very large scene sizes. Usually, the height and width of VHR optical remote sensing videos are hundreds of times larger than those of normal videos. This leads to background subtraction for VHR optical remote sensing videos being more time consuming. In addition, the whole VHR optical remote sensing video frame is too large to directly use with deep neural networks, such as CNNs.

To deal with the technical challenges mentioned above, in our article, we propose a deep learning framework that fully utilizes generative adversarial networks (GAN) and domain adaptation, which is capable of achieving reliable and efficient real-time background subtraction for VHR optical remote sensing videos. Specifically, we design a novel and robust background subtraction approach anchored on conditional generative adversarial networks (CGAN) [13] and domain-adversarial neural networks (DANN) [14]. By integrating CGAN with DANN, we evaluate and minimize the conflict between feature distributions of natural videos and VHR optical remote sensing videos so that the background subtraction performance for VHR optical remote sensing videos improves significantly.

The contributions of our article can be depicted as follows.  
- We choose the widely used dataset CDnet 2014 [15] as our natural videos and make a novel VHR optical remote sensing video dataset with no annotations. Excessive experiments are conducted on them. Numerous experiments demonstrate that our proposed approach is capable of achieving an average F-measure of 0.8533.
- We construct and train CGAN to learn a nonlinear mapping between the input video frame and generated background using natural videos. By leveraging CGAN, the proposed deep neural network has stronger generalization and more powerful feature extraction capability than conventional CNN and can achieve excellent background subtraction performance.
- DANN learns the distinction between the source domain (natural videos) and the target domain (VHR optical remote sensing videos). This means that the background subtraction method learned from the labeled natural videos can be performed on completely unlabeled VHR optical remote sensing videos. Benefiting from this, we greatly improve the background subtraction performance for VHR optical remote sensing videos compared with other advanced methods.
- We divide all of the VHR optical remote sensing video frames into regular blocks of the same size. Then, they are sent to the deep neural network in parallel and computed on the GPU. Finally, the outputs of the deep neural network are reverted to the original size, which is the background subtraction result. These steps make considerable contributions to address the problem of large scene size and significantly speed up processing.
- We propose a novel background subtraction method anchored on deep auto-encoder networks. Xu et al. [18] developed a neoteric background subtraction method anchored on deep auto-encoder networks. They first utilized a deep auto-encoder to extract dynamic background images from videos. Then, they leveraged another auto-encoder to learn the representation of the dynamic background. Kyungsun Lim et al. [19] designed a CNN with the structure of an encoder-decoder in which the encoder outputs a high-dimensional feature vector and the decoder converts it into the segmentation result.

Since CNN has achieved great success in image classification tasks, its application in the community of computer vision has become more and more extensive. Therefore, lots of novel methods for background subtraction are proposed based on CNN. Braham and Droogenbroeck [20] proposed a novel background subtraction method anchored
on spatial features learned from CNN. Cinelli [21] proposed a novel approach based on residual neural networks, which is able to detect moving objects by pixelwise foreground segmentation. Wang et al. [22] proposed a semiautomatic end-to-end network anchored on a multiresolution cascaded CNN. Babaei et al. [23] performed background subtraction from video sequences using CNN. The final background subtraction results were output after spatial-median filtering. Zhao et al. [24] proposed a two-stage CNN to generate background and foreground, which simultaneously output the background and foreground.

B. DOMA域 ADAPTATION
Domain adaptation holds the potential to transfer learned knowledge from the source domain to the discriminative but relevant target domain. In recent years, various methods of domain adaptation have been proposed using unsupervised learning. Huang et al. [25] presented a non-parametric method and directly produced resampling weights without distribution estimation, which matches distributions in high-dimensional feature space. Gong et al. [26] developed a novel approach that uses groundtruth to automatically bridge different distributions, which can optimize the original method discriminatively without any groundtruth from the target domain. Pan et al. [27] proposed transfer component analysis (TCA), a neoteric representation method designed specifically for domain adaptation. It projects input data onto the learned TCA, which can alleviate the discrepancy between two similar but different distributions.

While the above are unsupervised domain adaptation methods, there are multiple methods performing supervised domain adaptation by exploiting labeled data from the target domain. Gopalan et al. [28] were motivated by incremental learning so they designed intermediate representations of data between the source domain and the target domain, which holds the potential to learn a target domain discriminative classifier from projections of labeled source domain data. Baktashmotlagh et al. [29] proposed unsupervised domain adaptation, which was a novel type of domain invariant projection. It extracts invariant information across the source domain and the target domain. Specifically, it learns the projection of data from the low-level underlying space where the distance between the source domain and the target domain is minimized.

Existing domain adaptation approaches mainly focus on learning invariant feature representations across the source domain and the target domain. They seldom leverage low-level intrinsic structures. Gong et al. [30] developed a geodesic flow kernel, a novel type of kernel-based method, which aggregates infinite subspaces that are discriminative both in geometrical and statistical attributes from the source domain to the target domain. Chopra et al. [31] developed a neoteric model for domain adaptation anchored on deep learning. This model utilizes information among the source domain and target domain, which is capable of learning intermediate representations among the source domain and target domain.

C. GAN
Kocaoglu et al. [32] developed a generative model for a given causal graph by tapping into the advantages of the adversarial training process. They designed CausalGAN and CausalBEGAN, two novel types of conditional GAN. Experimental results demonstrate that the designed approach holds the potential to exploit underlying and volatile image distributions since the datasets are inevitably volatile. Feizi et al. [33] designed a novel GAN architecture to recover the maximum-likelihood solution and showed fast generalization capacity.

Hong et al. [34] developed a neoteric approach to aggregate GAN into fully convolutional networks (FCN), which can mitigate the gap between the source domain and the target domain. It utilizes a conditional generator and transforms features of synthetic images to real images. Then, a discriminator is utilized to distinguish them. In every training batch, the generator and the discriminator of CGAN compete against each other. By doing so, the generator and the discriminator of CGAN improve their performance.

III. METHODOLOGY
The proposed background subtraction framework for VHR optical remote sensing videos consists of four parts: (1) background image generation; (2) data preparation; (3) background subtraction network construction and training; (4) domain adaptation. Figure 1 depicts the framework of the background subtraction method for VHR optical remote sensing videos.

A. BACKGROUND GENERATION
For background subtraction using the proposed framework, we generate the background model from the original videos. The clear and vivid background model is vital for the proposed framework to fully learn the discrepancies between the current input frame and background model, which benefits for final background subtraction results.

We leverage the background generation method based on a foreground mask extracted by the SuBSENSE algorithm [35]. Figure 2 illustrates the generated background images extracted under highway and bungalow scenes. As shown in Figure 2, extracted background images are very similar to the real background. Although there are many moving targets in the scene, the generated background images are not blurred.

B. DATA PREPARATION
The RGB [36] and the HSV [37] are two usually utilized color spaces. Compared with RGB space, HSV space is more capable of showing visually brightness and darkness changes. In background subtraction, illumination variation, which has a negative effect on the performance of background subtraction, inevitably occurs in video scenes. For the purpose...
W. Yu et al.: Background Subtraction Based on GAN and Domain Adaptation for VHR Optical Remote Sensing Videos

In addition, with the purpose of improving the generalization ability of the proposed framework and allowing it to adapt to different scenes under various conditions, it is necessary to learn different sizes of videos in different scenes. In this article, a bilinear interpolation algorithm [38] is utilized to resize the original video frames in the CDnet 2014 dataset to the same size as $321 \times 321 \times 6$.

C. BACKGROUND SUBTRACTION NETWORK CONSTRUCTION AND TRAINING
1) BACKGROUND SUBTRACTION NETWORK FOR NATURAL VIDEOS BASED ON CGAN
To circumvent the problem that conventional GANs are unconstrained, Mehdi Mirza et al. proposed CGAN in 2014 by adding additional information, which is an extension of conventional GAN. A conditional variable $y$ with constraints is added to the generator and the discriminator to guide and restrain the generation process. The constraints can be a variety of information, such as category labels, partial data for image repair, and data from different modalities. When using category labels as constraints, the unsupervised GAN becomes supervised. This improvement has proven to be very effective, so CGAN has been widely used.

For the purpose of overcoming the shortcomings of conventional CNN, such as pool generalization and feature extraction capabilities, Long et al. proposed FCN [39] and successfully applied it to image semantic segmentation. FCN removes all fully connected layers in the conventional CNN. Transposed convolution layers in FCN restore the size of the feature map sampled by the pooled layer, which is designed to produce a correspondingly sized output for any size input. FCN not only preserves the overall structure information of the input image but also realizes pixel-by-pixel semantic segmentation.

The proposed background subtraction network based on CGAN is made up of a generator and discriminator. The generator is utilized to obtain the background subtraction result. It is modified from FCN. The discriminator is constructed by conventional CNN. It is utilized to discriminate the probability of the input image coming from the dataset or generator. Figure 3 shows the framework of the background subtraction network based on CGAN, which is called CGAN-BS. In Figure 3, the input of the generator consists of $z$ and $y$, where $z$ is composed of the original video frame and generated background and $y$ is the ground truth.

The backbone of the FCN in the generator is based on ResNet-50 [40], which is shown in Table 2. Figure 4 is the architecture of the generator. Table 2 gives the specific network parameters. In the FCN, the residual network is used to extract deep features hidden behind the original input image.

As illustrated in Figure 5, in the discriminator, the size of the input video frame is reduced through several convolutional layers with the stride of 2, ended by a fully connected layer. The output of the discriminator indicates the probability that the input video frame is coming from the generator or the dataset.
A mixed loss function is utilized to define the loss of the CGAN-BS. The losses of the generator and the discriminator are as follows:

$$G_{loss} = \frac{1}{N} \sum_{n=1}^{N} l_{bce}(s(x_n), y_n)$$

(4)

$$D_{loss} = \frac{1}{N} \sum_{n=1}^{N} l_{bce}(a(x_n, y_n), 1) + l_{bce}(a(x_n, s(x_n)), 0)$$

(5)

where $l_{bce}$ is the cross-entropy loss, $s(\cdot)$ is the generated result from the generator, $x$ is the training sample, $y$ is the groundtruth, $a$ is the output of the discriminator, $a(x_n, y_n)$ is the output of the discriminator when its input is the groundtruth, $a(x_n, s(x_n))$ is the output of the discriminator when its input is the output of the generator, and $n$ is the $n_{th}$ sample.

Assuming the predicted value is $z'$ and the actual value is $z$, the cross entropy is:

$$l_{bce}(z', z) = - [z \log(z') + (1 - z) \log(1 - z')]$$

(6)

During the training phase of the proposed model, we fix the generator and update the parameters of the discriminator, alternating iteratively. For the discriminator, it is expected that the discriminator can correctly distinguish whether the input video frame comes from the dataset or the generator. As a result, we minimize $D_{loss}$. For the generator, on the one hand, we expect that the generated result will be infinitely close to the groundtruth, that is, minimize $G_{loss}$. On the other hand, we expect that the generated result can "trick" the discriminator, that is, minimize $l_{bce}(a(x_n, s(x_n)), 1)$. Hence, we introduce the penalty factor $\lambda$ ($0 < \lambda < 1$) and modify the loss function of the generator to:

$$G_{mix} = G_{loss} + \lambda \cdot l_{bce}(a(x_n, s(x_n)), 1)$$

(7)

When training the generator, the parameters of the generator and the discriminator are updated according to $G_{mix}$. When the discriminator is trained, the parameters of the discriminator are updated according to $D_{loss}$.

2) BACKGROUND SUBTRACTION NETWORK FOR VHR OPTICAL REMOTE SENSING VIDEOS USING DANN

After we exploit CDnet 2014 to train background subtraction based on CGAN, the trained generator, which can be utilized to generate background subtraction results, is the desired result. We modify the trained generator by DANN, which is called DANN-BS.

DANN is a novel feature representation method designed for domain adaptation that concentrates on incorporating domain adaptation and deep feature learning at the same time. It is a generic approach that allows almost every existing feedforward deep neural network, including our proposed
generator, to implement knowledge transferability. The only additional component is the gradient reversal layer, which is utilized to reverse the gradient during the backpropagation computation process. DANN is made up of a label predictor, domain classifier and feature extractor.

In this article, we consider the background subtraction task as a classification task, which is aimed to classify the background and foreground for every pixel in the video frame. We define \( X \) as the input space and \( Y = \{0, 1, \ldots, L - 1\} \) as the set of \( L \) possible labels. Therefore, there are two different distributions over \( X \times Y \), which are named by source domain \( D_S \) and target domain \( D_T \). For unsupervised domain adaptation learning methods, the labeled source sample \( S \) are drawn from \( D_S \) while the unlabeled target sample \( T \) are drawn from \( D_X^T \), where \( D_X^T \) is the marginal distribution of \( D_T \) over \( X \). \( S \) and \( D \) both subject to independent and identical distribution. The domain adaptation algorithm can be presented as:

\[
S = \{(x_i, y_i)\}_{i=1}^n \sim (D_S)^n; \quad T = \{x_i\}_{i=n+1}^N \sim \left(D_X^T\right)^{n'},
\]

(8)

where \( N = n + n' \) is the total number of samples.

The target of domain adaptation is to create classifier \( \eta : X \rightarrow Y \) in a low target risk

\[
R(\eta) = \mathbb{P}_{(x,y) \sim D_T}(\eta(x) \neq y).
\]

(9)

In DANN-BS, it is assumed that there are two data distributions: source domain \( S(x, y) \) and target domain \( T(x, y) \). The goal of DANN-BS is to obtain better performance on the target domain using unsupervised learning. The implicit assumption is that the training samples are \( x_1, x_2, \ldots, x_N \) from the source domain \( S(x, y) \) and the target domain \( T(x, y) \). At the same time, we define the domain label of the \( i \)th training sample as \( d_i \), where \( d_i \in \{0, 1\} \). If \( d_i = 0 \), it means \( x_i \sim S(x) \). If \( d_i = 1 \), it means \( x_i \sim T(x) \).

In the course of training, the inputs of DANN-BS come from source domain datasets with labels and target domain datasets without labels, as well as domain classification labels for the source domain and the target domain. Namely, we have the background subtraction ground truth for the source domain but do not have the background subtraction ground truth for the target domain.

Since the VHR optical remote sensing video dataset is not labeled, there is no groundtruth. As a result, the input of the generator is changed from \( 321 \times 321 \times 7 \) to \( 321 \times 321 \times 6 \). Figure 6 shows the framework of the background subtraction network using DANN. As visualized in Figure 6, the green dotted box is the feature extractor, the blue dotted box is the label predictor and the red dotted box is the domain classifier.

The connecting line between the feature extractor and domain classifier is the gradient reversal layer. When performing forward propagation, the gradient reversal layer is
identified as an identity connection. When performing back-
propagation, a negative constant is multiplied by the original
gradient in terms of the gradient reversal layer. The mathe-
matical expression is as follows:

\[ R_\lambda(x) = x \]  \hspace{1cm} (10)

\[ \frac{dR_\lambda}{dx} = -\lambda I \]  \hspace{1cm} (11)

The input of the feature extractor is a batch of 321 \times 321 \times 6
tensors. The input batch has two parts: the first part comes
from natural videos, and the second part comes from VHR
optical remote sensing videos. The whole batch is fed into the
feature extractor, and we obtain the corresponding features.

The output of the label predictor is the background sub-
traction. It is worth mentioning that in the course of train-
ing, the input of the label predictor is the first part of the
batch, which means that the second batch coming from VHR
optical remote sensing videos is discarded. The reason is
that the VHR optical remote sensing video dataset has no
groundtruth, so backpropagation cannot be conducted for
the second batch.

In the domain classifier, there is a fully connected layer and
two convolution layers. The input of the domain classifier is
a complete batch including natural videos and VHR optical
remote sensing videos. The outputs of the domain classifier
determine whether the input batch comes from the source
domain (natural videos) or the target domain (VHR optical
remote sensing videos).

In DANN-BS, the feature extractor is aimed to learn a hid-
den representation either for source frames or target frames
so that the domain classifier is able to classify source domain
inputs and target domain inputs accurately and the label
predictor is able to classify the source inputs.

In the course of training, the learning rate is adjusted using
the following formula:

\[ \mu_p = \frac{\mu_0}{(1 + \alpha \cdot p)^\beta} \]  \hspace{1cm} (12)

where \( p \) is gradually increasing from 0 to 1, \( \mu_0 = 0.01, \alpha = 9 \)
and \( \beta = 0.69 \). It can facilitate convergence and alleviate
deviation in the target domain. In addition, the momentum
is set to 0.9.

The \( \lambda_p \) is set to 0 during the initialization phase. Then,
it linearly increases to 1, as in the formula below:

\[ \lambda_p = \frac{2}{1 + \exp(-\gamma \cdot p)} - 1 \]  \hspace{1cm} (13)

where \( \gamma \) is set to 10. It makes the domain classifier immune to
noise during the initialization phase of training. \( \lambda_p \) is utilized
to update the feature extractor.

IV. EXPERIMENTAL RESULTS AND ANALYSIS
We conducted multiple experiments on the CDnet 2014
dataset and VHR optical remote sensing video dataset, which
we made, to evaluate the performance of our proposed back-
ground subtraction method.

A. DATASET
CDnet 2014 is one of the benchmark datasets for background
subtraction, which includes some complex video scenes that
are challenging in the field of background subtraction, such as
camera jitters, dynamic backgrounds, shadows, and obscure
videos.

Our dataset is based on the VHR optical remote sensing
videos provided by the Jilin No. 1 satellite. The videos are
in AVI format, and the resolution of videos is less than 1m.
The ground area covered by each video is 11km \times 4.5km.
All videos have undergone geometric correction, radiometric
correction and image stabilization. The targets in our VHR
optical remote sensing videos are various. The average length
of each video is 30 seconds. Figure 7 shows one of the VHR
optical remote sensing video frames in Beijing.

A local binary pattern (LBP) [41] texture histogram is uti-
lized to demonstrate that the two datasets have scrupulously
different distributions. LBP is a popular local texture descrip-
tor in computer vision that has also been widely applied in
background subtraction [42]–[44]. Hence, LBP texture his-
tograms are of great value to help us observe the distributions
of two datasets.
As shown in Figure 8, (a) is the LBP texture histogram of CDnet 2014 and (b) is the LBP texture histogram of the VHR optical remote sensing video dataset. They both reflect the overall local texture feature of video frames in two datasets.
In (a), only the bin of 255 has the greatest number of pixels. However, in (b), bin 25 and bin 225 have many pixels in addition to bin 255.

As a result, we, therefore, find it prudent to comment that the CDnet 2014 and VHR optical remote sensing video datasets share similar but different distributions.

Since the resolution of optical remote sensing videos was 12,000 × 5,000 and include considerable interference, such as background clutter and moving cloud occlusion, we processed the videos first. The whole video frame was cut into extensive blocks. Each block had the same size of 321 × 321. Then, they were sent to the background subtraction network in parallel and computed on the GPU. Finally, all the outputs of the background subtraction network were reverted to the original size, which was the background subtraction result. These steps made considerable contributions to address the problem of large scene size and significantly speed up processing.

B. EVALUATION METRICS

We note that the built-in assumption of existing proposals is that background subtraction can be regarded as a pixel-level classification issue. Its essence is the two-class classification of image pixels. Therefore, evaluation metrics based on pixel-level classification were used to evaluate the proposed method in this article and other advanced methods.

In this article, we treat background pixels as positive while foreground pixels as negative. Pixel-level classification results can be divided into four categories: background pixels are judged as background pixels, foreground pixels are judged as background pixels, foreground pixels are judged as foreground pixels, and background pixels are judged as foreground pixels, represented by true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

In terms of the above statistical indicators, seven objective evaluation indicators including Recall (Re), Specificity (Sp), Precision (Pr), F-Measure (FM), False Positive Rate (FPR), False Negative Rate (FNR) and Percentage of Wrong Classifications (PWC) were used to evaluate the background subtraction, which are presented in this article vividly [45].

In background subtraction, Re denotes that the ratio of pixels that should be judged as the background is correctly classified. The larger the value of Re, the better the performance. Pr represents the ratio of pixels that are precisely classified among the pixels judged as background. The larger the value of Pr, the better the performance. Sp mainly characterizes the ability to classify foreground pixels. The larger the value of Sp, the better the performance. FPR, FNR and PWC all indicate the ratio of misclassifications. The smaller the value is, the better the performance. In practice, Re and Pr are usually used to measure the background subtraction. Re and Pr are expected to be as high as possible. However, in fact, they are contradictory in some cases. It is possible for Re to high but Pr to be low. Therefore, in general, the comprehensive indicator FM is generally used to evaluate background subtraction.

C. EXPERIMENTAL ENVIRONMENT

Our hardware experimental environment: The CPU was an Intel Xeon (R) E5-2640 v4 with 20 cores running at 2.4 GHz. The memory size was 64 GB. The GPU was an NVIDIA GeForce GTX 1080 with 8 GB of memory.

D. FEASIBILITY EXPERIMENT AND ANALYSIS

To verify the feasibility of the proposed DANN-BS, the training time and inference time were analyzed.

1) TRAINING TIME OF THE PROPOSED METHOD

When batch size was set to 30, the training time was tested under the CPU and GPU. With the CPU, the training time was more than one month, and the memory consumption was more than 20 GB. With GPU acceleration technology, the training time was reduced to 18 hours, which was within the acceptable range.

2) INFEERENCE TIME OF THE PROPOSED METHOD

The trained model was utilized for background subtraction. Experiments show that the single image inference time is 50 ms, or 20 frames per second. As a result, the proposed method was capable of being processed in real-time.

E. EXPERIMENTS ON BACKGROUND SUBTRACTION NETWORK CGAN-BS

To further illustrate the superiority of the proposed network over existing methods, several representative background subtraction algorithms were selected for comparison. Background subtraction algorithms based on unsupervised learning include the mixed Gaussian model (GMM) [47], the pixel-based adaptive segmentation PBAS algorithm [48] and the SuBSENSE algorithm [35]. The background subtraction algorithm based on supervised learning was the DeepBS algorithm proposed by Babaee et al. in 2017 [23].

1) EXPERIMENTS ON CDnet 2014

We leverage the CDnet 2014 dataset to train the proposed background subtraction network CGAN-BS. Figure 9 is a diagram showing the background subtraction results of the 1,500th frame, the 1,840th frame, the 1,900th frame, and the 2,360th frame in the ‘fall’ video scene by the proposed CGAN-BS and other advanced methods. The ‘fall’ video scene belongs to the dynamic background category in the CDnet 2014 dataset. Due to the shaking of leaves in the ‘fall’ video scene, the traditional unsupervised learning method incorrectly classified the swaying leaves as foreground. In addition, there was considerable noise around the foreground. The DeepBS method based on a deep neural network accurately classified the swaying leaves into backgrounds. However, there were too many holes in the foreground, as shown in the classification results. In the 1,900th frame of the ‘fall’ video, other advanced methods classified the suspected regions into the foreground incorrectly. The foreground classified by the proposed CGAN-BS method...
was consistent, and the details of the foreground along with the structural information were complete. More importantly, the suspect regions were not classified as foreground by the proposed CGAN-BS. Consequently, the proposed CGAN-BS surpassed other methods.

Table 3 shows evaluation indicators of CGAN-BS compared with other advanced methods in the CDnet 2014 dataset. By leveraging CGAN, the proposed background subtraction network CGAN-BS has stronger generalization and more powerful feature extraction capability than conventional CNN and other advanced methods, which obtained excellent background subtraction performance.

2) EXPERIMENTS ON THE VHR OPTICAL REMOTE SENSING VIDEO DATASET

Taking a VHR optical remote sensing image block of Beijing as an example, we carried out experiments on it and obtained the background subtraction results. Figure 10 is a diagram showing the background subtraction results of CGAN-BS and other advanced methods in the VHR optical remote sensing video.

From Figure 9, Figure 10 and Table 4, we can see that the proposed background subtraction network CGAN-BS shows good performance on natural videos while showing poor performance on the VHR optical remote sensing video.
video dataset. It shows that they share similar but different distributions.

In Table 4, the experimental consequences of CGAN-BS in untrained VHR optical remote sensing video datasets were
TABLE 4. Average Evaluation indicators of CGAN-BS using two datasets.

| Dataset                               | Re   | Sp   | FPR  | FNR  | PWC  | Pr   | FM  |
|---------------------------------------|------|------|------|------|------|------|-----|
| CDnet 2014                            | 0.8350 | 0.9863 | 0.0138 | 0.1650 | 0.0440 | 0.9382 | 0.8836 |
| VHR optical remote sensing videos     | 0.7800 | 0.9500 | 0.0413 | 0.2200 | 0.0770 | 0.8254 | 0.8021 |

TABLE 5. Average Evaluation indicators of DANN-BS in VHR optical remote sensing videos dataset.

| Evaluation indicators | DANN-BS | CGAN-BS | PBAS | SuBSENSE | DeepBS | GMM |
|-----------------------|---------|---------|------|----------|--------|-----|
| FM                    | 0.8533  | 0.8021  | 0.7411 | 0.7360 | 0.7815 | 0.7577 |
| Pr                    | 0.9345  | 0.8254  | 0.7526 | 0.7474 | 0.8042 | 0.7614 |
| Re                    | 0.7850  | 0.7800  | 0.7300 | 0.7250 | 0.7600 | 0.7500 |

Experiments demonstrate that our proposed CGAN-BS obtains good performance in video scenes that have been involved in training, but it does not work well in untrained video scenes. Therefore, the robustness of the proposed CGAN-BS needs to be further improved.

F. EXPERIMENTS ON BACKGROUND SUBTRACTION NETWORK DANN-BS

1) EXPERIMENTAL RESULTS

We used the unlabeled VHR optical remote sensing video dataset that we made to test the proposed DANN-BS. It is not as good as those of trained natural videos. What is worse is that the FNR and PWC were high. This shows that the performance of the CGAN-BS on untrained video scenes is flawed.

Experiments demonstrate that our proposed CGAN-BS obtains good performance in video scenes that have been involved in training, but it does not work well in untrained video scenes. Therefore, the robustness of the proposed CGAN-BS needs to be further improved.

Figure 11 shows background subtraction results of CGAN-BS and DANN-BS in the VHR optical remote sensing videos. From Figure 11, we can see that the proposed DANN-BS can effectively improve the generalization ability of background subtraction for unlabeled VHR optical remote sensing videos and can also be applied to other unlabeled complex video scenes owing to domain adaptation.

VOLUME 8, 2020 119155
background subtraction network DANN-BS obviously improves the performance.

2) EXPERIMENTS WITH DIFFERENT COLOR SPACES

In the VHR optical remote sensing video dataset, when the input video frames use the RGB color space, the average FM value of 0.8533 is higher than that of 0.8509, where the input video frames use the HSV color space. Nonetheless, the differences between them were small. Consequently, the selection of different color spaces had little effect on the background subtraction results for the proposed DANN-BS method.

V. CONCLUSION

In this article, we designed a neoteric and outstanding background subtraction network via CGAN and DANN. The proposed method utilizes the generated background and video frame pairs and obtains the background subtraction result for VHR optical remote sensing videos by using domain adaptation. Our experiments on the CDnet 2014 dataset and our dataset demonstrated that our proposed method achieved an average FM of 0.8533, which reveals excellent performance on background subtraction for VHR optical remote sensing videos.

It is noteworthy that the quality and quantity of background subtraction datasets for VHR optical remote sensing videos in academia are limited. In the future, we will devote our effort to collecting more background subtraction datasets to support in-depth research.

REFERENCES

[1] Q. Shi, X. Liu, and X. Li, “Road detection from remote sensing images by generative adversarial networks,” *IEEE Access*, vol. 6, pp. 25486–25494, 2018.

[2] S. Yin, Y. Zhang, and S. Karim, “Large scale remote sensing image segmentation based on fuzzy region competition and Gaussian mixture model,” *IEEE Access*, vol. 6, pp. 26069–26080, 2018.

[3] Z. Zhang, R. Jiang, S. Mei, S. Zhang, and Y. Zhang, “Rotation-invariant feature learning for object detection in VHR optical remote sensing images by double-net,” *IEEE Access*, vol. 8, pp. 20818–20827, 2020.

[4] W. Li, H. Liu, Y. Wang, Z. Li, Y. Jia, and G. Gui, “Deep learning-based classification methods for remote sensing images in urban built-up areas,” *IEEE Access*, vol. 7, pp. 36274–36284, 2019.

[5] X. Li, Z. Li, R. Feng, S. Luo, C. Zhang, M. Jiang, and H. Shen, “Generating high-quality and high-resolution seamless satellite imagery for large-scale urban regions,” *Remote Sens.*, vol. 12, no. 1, p. 81, Dec. 2019.

[6] F. Bi, M. Lei, Y. Wang, and D. Huang, “Remote sensing target tracking in UAV aerial video based on saliency enhanced MNet,” *IEEE Access*, vol. 7, pp. 76731–76740, 2019.

[7] W. Ma, Q. Guo, Y. Wu, W. Zhao, X. Zhang, and L. Jiao, “A novel multi-model decision fusion network for object detection in remote sensing images,” *Remote Sens.*, vol. 11, no. 7, p. 737, Mar. 2019.

[8] B. Garcia-Garcia, T. Bouwmans, and A. J. Rosales Silva, “Background subtraction in real applications: Challenges, current models and future directions,” *Comput. Sci. Rev.*, vol. 35, Feb. 2020, Art. no. 100204.

[9] S. Javed, A. Mahmood, T. Bouwmans, and S. K. Jung, “Background–foreground modeling based on spatiotemporal sparse subspace clustering,” *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5840–5854, Dec. 2017.

[10] F. El Baf, T. Bouwmans, and B. Vachon, “Fuzzy integral for moving object detection,” in *Proc. IEEE Int. Conf. Fuzzy Syst. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1729–1736.

[11] M. Braham, S. Pierard, and M. Van Droogenbroeck, “Semantic background subtraction,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4552–4556.

[12] D. Sakkos, H. Liu, J. Han, and L. Shao, “End-to-end video background subtraction with 3D convolutional neural networks,” *Multimedia Tools Appl.*, vol. 77, no. 17, pp. 23023–23041, Sep. 2018.

[13] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” 2014, arXiv:1411.1784. [Online]. Available: http://arxiv.org/abs/1411.1784

[14] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 2096, 2016.

[15] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, “CDNet 2014: An expanded change detection benchmark dataset,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 387–394.

[16] R. Guo and H. Qi, “Partially-sparse restricted Boltzmann machine for background modeling and subtraction,” in *Proc. 12th Int. Conf. Mach. Learn. Appl.*, vol. 1, Dec. 2013, pp. 209–214.

[17] L. Xu, Y. Li, Y. Wang, and E. Chen, “Temporally adaptive restricted Boltzmann machine for background modeling,” in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1938–1944.

[18] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann, “Deep learning for background subtraction with 3D convolutional neural networks,” 2014, arXiv:1411.1784.

[19] K. Lim, W.-D. Jiang, and C.-S. Kim, “Background subtraction using encoder-decoder structured convolutional neural network,” in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Survell. (AVSS)*, Aug. 2017, pp. 1–6.

[20] M. Braham and M. Van Droogenbroeck, “Deep background subtraction with scene-specific convolutional neural networks,” in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, May 2016, pp. 1–4.

[21] L. P. Cinelli, “Anomaly detection in surveillance videos using deep residual networks,” Ph.D. dissertation, Dept. Electron. Comput. Eng., Univ. Federal do Rio de Janeiro, Rio de Janeiro, Brazil, 2017.

[22] Y. Wang, Z. Luo, and P.-M. Jodoin, “Interactive deep learning method for segmenting moving objects,” *Pattern Recognit. Lett.*, vol. 96, pp. 66–75, Sep. 2017.

[23] M. Babaei, D. T. Dinh, and G. Rigoll, “A deep convolutional neural network for video sequence background subtraction,” *Pattern Recognit.*, vol. 76, pp. 635–649, Apr. 2018.

[24] X. Zhao, Y. Chen, M. Tang, and J. Wang, “Joint background reconstruction and foreground segmentation via a two-stage convolutional neural network,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 343–348.

[25] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. J. Smola, “Correcting sample selection bias by unlabeled data,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 601–608.

[26] B. Gopalan, R. Krause, and F. Sha, “Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation,” in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 222–230.

[27] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[28] R. Gopalan, R. Li, and R. Chellappa, “Domain adaptation for object recognition: An unsupervised approach,” in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 999–1006.

[29] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann, “Unsupervised domain adaptation by domain invariant projection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 769–776.

[30] B. Gopalan, Y. Shi, F. Sha, and K. Grauman, “Geodesic flow kernel for unsupervised domain adaptation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2012, pp. 2066–2073.

[31] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[32] R. Gopalan, R. Li, and R. Chellappa, “Domain adaptation for object recognition: An unsupervised approach,” in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 999–1006.

[33] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann, “Unsupervised domain adaptation by domain invariant projection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 769–776.
[35] P.-L. St-Charles, G. A. Bilodeau, and R. Bergevin, “Subsense: A universal change detection method with local adaptive sensitivity,” IEEE Trans. Image Process., vol. 24, no. 1, pp. 359–373, Jan. 2015.

[36] M.-H. Shi, L. Shen, S.-Z. Long, and X.-M. Hu, “The revision of conversion formula from RGB color space to HSV color space,” Basic Sci. J. Text Univ., vol. 21, no. 3, pp. 351–356, 2014.

[37] S. Sural, G. Qian, and S. Pramanik, “Segmentation and histogram generation using the HSV color space for image retrieval,” in Proc. Int. Conf. Image Process., vol. 2, Sep. 2002, p. II.

[38] E. J. Kirkland, “Bilinear interpolation,” in Advanced Computing in Electron Microscopy, Boston, MA, USA: Springer, 2010, pp. 261–263.

[39] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431–3440.

[40] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[41] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” Pattern Recognit., vol. 29, no. 1, pp. 51–59, Jan. 1996.

[42] A. Shimada and R.-I. Taniguchi, “Hybrid background model using spatial-temporal LBP,” in Proc. 6th IEEE Int. Conf. Adv. Video Signal Based Surveill., Sep. 2009, pp. 19–24.

[43] G. Xue, L. Song, J. Sun, and M. Wu, “Hybrid center-symmetric local pattern for dynamic background subtraction,” in Proc. IEEE Int. Conf. Multimedia Expo, Jul. 2011, pp. 1–6.

[44] T. W. Chua, Y. Wang, and K. Leman, “Adaptive texture-color based background subtraction for video surveillance,” in Proc. 19th IEEE Int. Conf. Image Process., Sep. 2012, pp. 1–6.

[45] M. Camplani, L. Maddalena, G. M. A. Alcover, A. Petrosino, and L. Salgado, “A benchmarking framework for background subtraction in RGBD videos,” in Proc. Int. Conf. Image Anal. Process., 2017, pp. 219–229.

[46] N. M. Oliver, B. Rosario, and A. P. Pentland, “A Bayesian computer vision system for modeling human interactions,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 831–843, 2000.

[47] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2, Jun. 1999, pp. 246–252.

[48] M. Hofmann, P. Tiefenbacher, and G. Rigoll, “Background segmentation with feedback: The pixel-based adaptive segmenter,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2012, pp. 38–43.

WENTAO YU (Student Member, IEEE) received the B.S. degree in electronic information engineering from the Chengdu University of Information Technology, Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree with Xidian University, Xi’an, China. His current research interests include deep learning and computer vision.

JING BAI (Member, IEEE) received the B.S. degree in electronic and information engineering from Zhengzhou University, in 2004, and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, in 2009. She is currently an Associate Professor with Xidian University. Her research interests include image processing, machine learning, and intelligent information processing.

LICHENG JIAO (Fellow, IEEE) received the B.S. degree from Shanghai Jiao Tong University, in 1982, and the M.S. and Ph.D. degrees from Xi’an Jiaotong University, Xi’an, China, in 1984 and 1990, respectively. He has been a Professor with the School of Artificial Intelligence, Xidian University, Xi’an, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education of China. His research interests include image processing, natural computation, machine learning, and intelligent information processing.

* * *