Abstract—Objective: Dim target detection in remote sensing images is a significant and challenging problem. In this work, we seek to explore event-related brain responses of dim target detection tasks and extend the brain-computer interface (BCI) systems to this task for efficiency enhancement. Methods: We develop a BCI paradigm named Asynchronous Visual Evoked Paradigm (AVEP), in which subjects are required to search the dim targets within satellite images when their scalp electroencephalography (EEG) signals are simultaneously recorded. In the paradigm, stimulus onset time and target onset time are asynchronous because subjects need enough time to confirm whether there are targets of interest in the presented serial images. We further propose a Domain adaptive and Channel-wise attention-based Time-domain Convolutional Neural Network (DC-tCNN) to solve the single-trial EEG classification problem for the AVEP task. In this model, we design a multi-scale CNN module combined with a channel-wise attention module to effectively extract features also illustrate the effectiveness and interpretability of our proposed paradigm and learning model. Conclusion: The proposed paradigm and model can effectively explore ambiguous event-related brain responses on EEG-based dim target detection tasks. Significance: Our work can provide a valuable reference for BCI-based image detection of dim targets.

Index Terms—Asynchronous visual evoked paradigm (AVEP), brain-computer interface (BCI), channel-wise attention, dim target detection, domain adaptation, electroencephalography (EEG).

I. INTRODUCTION

RECENTLY, brain-computer interface (BCI) has started to extend its applications from helping people with limitations in motor controls or communication to augmenting human capabilities [1]–[3], such as speeding up the process of finding targets of interest in large collections of images [4]–[6]. This BCI-based target-searching technique can typically be applied to many real-life applications, including counterintelligence, policing, and professional interpretation of images captured by drones or satellites with trained experts. In these studies, the rapid serial visual presentation (RSVP) protocol is commonly used to detect whether targets of interest are contained in the presented images by simultaneously recording electroencephalography (EEG) signals during the tasks [4]–[9].

The RSVP detection task is synchronous because conspicuous targets can be discovered once images are presented. However, in many realistic applications of image target search, such as satellite and high-altitude unmanned aerial vehicle image interpretation, targets of interest can hardly be found as rapidly as the presentation rates in the RSVP. Moreover, similar environmental backgrounds, different observing angles, illumination variations, and occlusions would significantly increase the difficulty of target detection. In these tasks, the subjects require sufficient time to search and identify the targets of interest. The occurrence of finding dim targets is unpredictable, and depends mainly on the subjects. In other words, the stimulus onset and target onset times are asynchronous. Therefore, a paradigm with a slower image presentation speed, called the Asynchronous Visual Evoked Paradigm (AVEP), is needed for dim target detection.

A longer search period in AVEP brings new challenges to the BCI-based detection of images with targets. First, owing to the larger noise appearing in the long-term EEG signals, the signal components would become complex, and the signal baseline drifts severely, bringing greater challenges to feature extraction [15]. Second, the target onset time is inevitably changeable across trials and subjects because it is difficult to control the moment at which observers find targets of interest in a typical search task. Since it is difficult to average multiple event-related potential (ERP) responses without the alignment of event onsets, single-trial ERP detection is important for dim target detection tasks. Finally, ERP signals have been shown to be associated with cognitive processing [6], [7], [10]. There could be a distribution discrepancy in ERP signals across
subjects due to individual differences in cognitive processing. Therefore, a BCI-based detection method needs to better deal with inter-individual differences.

In this paper, we propose a Domain adaptive and Channel-wise attention-based Time-domain Convolutional Neural Network (DC-tCNN) for BCI-based detection of dim image targets leveraging an AVEP paradigm. Given that multi-channel EEG signals contain some irrelevant and redundant information, we adapt the channel-wise attention mechanism to focus on more important channels by adaptively assigning attention weights. Since the feature distributions are diverse among individuals, we apply the domain adaptation to mitigate the cross-subject distribution discrepancy. We then integrate the channel-wise attention mechanism and domain adaptation in a unified framework with a multi-scale 1-D temporal convolutional neural network. We demonstrated that compared with the conventional classification strategy and typical EEG deep learning networks, the proposed DC-tCNN could yield superior performance for single-trial classification in the detection of dim targets. The major contributions of this study are as follows:

1. To detect dim targets in remote sensing images based on BCI, we develop the AVEP paradigm in which the target onset time may lag behind the stimulus onset time in a subject-driven manner. To the best of our knowledge, this is the first study to use dim targets of remote sensing images as stimuli for BCI-based single-trial image classification.

2. To address the problem of modeling long-term EEG signals and discrepancy of feature distributions across subjects, we propose a deep learning model combining multi-scale time-domain CNN, channel-wise attention, and domain adaptation modules (DC-tCNN). In this model, the multi-scale time-domain CNN aims to capture the underlying event-related brain responses in long-term EEG signals. The channel attention module can retrieve important attentive information from the multi-channel signals and reduce the interference of irrelevant signals. Moreover, domain adaptation can effectively mitigate the distribution discrepancies across subjects.

3. To evoke ERP signals in the BCI experiment, the number of target samples is much less than that of non-target samples. Therefore, limited target samples and class imbalance also increase the challenges of single-trial classification. We apply random sampling and an ensemble learning approach during the training period to enhance classification robustness. The experimental results demonstrate the superiority of the proposed model. Visualization analysis of spatiotemporal features also indicates their effectiveness and interpretability.

II. RELATED WORK

A. Image Target Detection Based on Brain-Computer Interfaces

Recently, RSVP-based BCI has been used to detect and recognize objects or other targets of interest [4], [6], [8], [11]–[13], which would benefit professionals by reducing the burden of reviewing many images daily. The process of RSVP is displaying image sequences in a fixed focal position at high presentation rates. The target onset is definite and synchronous with the stimulus onset because the targets can be found once the images are presented. Bigdely-Shamlo et al. [11] designed an RSVP experiment in which participants were asked to search for images containing airplanes in rapidly presented image clips. They reported high accuracies for single-trial classification based on 128-channel EEG data using independent component analysis (ICA). Matran-Fernandez et al. [6] also explored the possibility of using ERPs to extract information regarding the spatial location of targets in an RSVP experiment. They found a significant correlation between ERPs and the horizontal location of targets in the aerial images.

RSVP-based target detection generally uses image sequences with conspicuous targets. However, in this study, we use dim targets in satellite images as stimuli to elicit ERPs. The target onset time and stimulus onset time are asynchronous because of unpredictable detection latency. Similarly, Song et al. [20] designed a video target detection task in which participants were asked to detect vehicles in the video. The target stimulus may appear at any time of the video. They then developed an asynchronous detection framework by aligning signals on a template to solve this problem.

B. Single-Trial EEG Classification Methods

In previous studies, it is common for single-trial EEG classification to be achieved by using several feature extraction algorithms based on a two-stage machine learning framework. Since the observed brain responses are signals with low signal-to-noise ratio (SNR), many algorithms have been proposed for noise reduction and feature extraction in the single-trial classification task. Cecotti et al. [12] used xDAWN and common spatial pattern to improve SNR. The filtered signals were then sent to linear classifiers, such as Bayesian linear discriminant analysis and support vector machines (SVM), for single-trial ERP detection. Moreover, ICA was used to extract time and time-frequency domain features. Subsequently, single-trial classification was achieved using Fisher discriminative classifiers [11]. In addition, hierarchical discriminant component analysis was developed for single-trial analysis [4], [14].

With the development of deep learning, deep neural networks have been widely used in single-trial EEG classification. The convolutional neural network, as an end-to-end deep learning model, is utilized to extract spatiotemporal features from EEG data [15], [16]. Recently, several CNN variations have been proposed to enhance the accuracy of single-trial EEG classification. For example, EEGNet with depth-wise and separable convolution was proposed to extract features more efficiently [17]. Zang et al. [18] designed a novel model combining a standard convolutional layer, a permute layer, and a depth-wise convolution layer. The results indicated that the model could make full use of the phase-locked characteristic and achieve better performance in single-trial EEG classification. Li et al. [8] proposed a phase preservation neural network consisting of dilated temporal convolution layers, a spatial convolution layer, and a fully connected layer. This model could improve EEG classification by considering the phase information of the ERPs.
C. Channel-Wise Attention and Domain Adaptation

Pioneered by attention perception, channel-wise attention is designed to adaptively re-weight the channels, which can retrieve more important information from multi-channel signals. Recently, channel-wise attention has been integrated into neural network architectures [19]–[22]. Woo et al. [22] proposed a convolutional block attention module and integrated the module into CNN architectures to improve the classification and detection performances. Meanwhile, Chen et al. [19] incorporated spatial and channel-wise attention into a CNN for image captioning. The model outperformed state-of-the-art methods. Tao et al. [21] also integrated channel-wise attention into a CNN to explore more discriminative features for EEG emotion recognition. The experimental results showed that the weights of channels related to emotions were greater than the others. Lan et al. [20] applied a multi-attention mechanism to concentrate on the important channels and discriminative temporal periods.

The significant discrepancy in feature distributions caused by individual differences limits the generalizability of EEG signals classifiers [23]. Therefore, applying the model across individuals is usually challenging in single-trial EEG classification. Recently, domain adaptation methods have been proposed to reduce the discrepancy in feature distribution across domains [24]–[28]. Long et al. [29] proposed the deep adaptation network to learn transferable features for domain adaptation without considering conditional information. Furthermore, marginal and conditional distributions were adopted for joint distribution adaptation [30]. In particular, adversarial learning is embedded in domain adaptation. Li et al. [25] proposed adapting the joint distribution with adversarial training to mitigate discrepancies in latent representations for EEG emotion recognition. Since the EEG data are often collected from different subjects, the source domain can be regarded as a multi-source domain. Wei et al. [13] developed a multi-source conditional adversarial domain adaptation framework to further improve model performance by integrating multiple domain adaptation results.

III. PROPOSED ALGORITHM

A. Model Architecture Overview

Here we introduce the proposed DC-tCNN model, including the channel attention module, multi-scale convolutional module, and domain adaptation module. As shown in Fig. 1, the model is divided into two parts. The first part is the channel attention and multi-scale convolutional modules (Fig. 1A), which are used as the encoder to extract features from EEG samples. The second part is the domain adaptation using an adversarial network (Fig. 1B) that aims to align the feature distributions between subjects.

In this study, since there are only a few target samples in one session for a single subject, we compare our model with state-of-the-art methods only for cross-subject classification tasks using a leave-one-subject-out method. In the cross-subject classification task, the training set is regarded as the source domain \( \mathcal{D}_s = \{ (x^s_i, y^s_i) \}_{i=1}^{n_s} \) of \( n_s \) labeled samples, and the testing set is regarded as the target domain \( \mathcal{D}_t = \{ x^t_j \}_{j=1}^{n_t} \) of \( n_t \) unlabeled samples. The joint distributions of source and target domains are \( P(x^s, y^s) \) and \( Q(x^t, y^t) \) \((P \neq Q)\). The training batches consist of the same numbers of source and target samples. During training, we use not only the source label predictor \( g = G(x) \) for the classification task, but also the domain predictor \( d = D(x) \) to mitigate the cross-domain distribution discrepancy \( disc(P, Q) \) to optimize the model parameters. Note that the different feature distributions are matched by conditioning feature representation \( f \) and label prediction \( g \) [31], [32]. In other words, the model simultaneously minimizes the source single-trial classification error and discrepancy of the feature distributions. The hidden features from both domains are fed into the domain predictor, which is a binary classifier that distinguishes the source domain from the target domain. Meanwhile, the hidden features from the two domains are fed into the label predictor to output the task labels. During test, samples of the target domain are fed into the model only.

B. Channel-Wise Attention Module

Recent studies have focused on the attention mechanism [19], [21], which is often used to model the dependencies of sequences without considering their distance. Moreover, the attention mechanism can also be used to select the most contributed electrodes, which are highly associated with the target tasks. In this study, to explore the importance of different channels and reduce the interference of irrelevant information, we apply a channel-wise attention module to adaptively assign the weights of each channel by exploiting the inter-channel relationship of the features.

Let \( x_i \in \mathbb{R}^{C \times L} \) represent an input EEG sample, where \( C \) denotes the number of channels and \( L \) denotes the number of sampling points. The channel-wise attention module aims to infer a 1D channel weight map \( v_i \in \mathbb{R}^{C \times 1} \). The detailed
procedure is described as follows. First, the mean pooling operation is applied for all preprocessed EEG samples \( \{x_i\}_{i=1}^{n} \), generating the mean-pooled features of each channel \( \hat{x}_i = [\hat{x}_{i,1}, \hat{x}_{i,2}, \cdots, \hat{x}_{i,j}, \cdots, \hat{x}_{i,C}] \). Here, \( j = 1, 2, \cdots, C \) denotes the \( j \text{th} \) channel. The mean-pooled features are then fed into two fully connected layers to produce a channel weight map, which can be formulated in the following form:

\[
v_i = \text{softmax} (w_2 \cdot (\tanh (w_1 \cdot \hat{x}_i + b_1)) + b_2)
\]

(1)

where \( w_1, w_2, b_1, b_2 \) are trainable weights (\( w \)) and biases (\( b \)), respectively. \( \tanh \) and \( \text{softmax} \) denote \( \tanh \) activation function and \( \text{softmax} \) function, respectively. The channel weight map \( v_i \) represents the weights of each channel for \( x_i \).

Then, \( v_i \in \mathbb{R}^{C \times L} \) is obtained by broadcasting the channel weight map \( v_i \) along the time dimension. Finally, the extracted EEG features are generated by multiplying \( x_i \) with \( v_i \):

\[
\hat{x}_i = x_i \otimes v_i
\]

(2)

where \( \otimes \) represents element-wise multiplication.

C. Multi-Scale Convolutional Module

The onset of event-related brain responses in this experiment varies across trials and subjects. Previous studies have also found that there are individual differences in the waveforms of EPRs [33], [34]. Therefore, a fixed length of time domain convolutional layer may not extract sufficient information because it can only capture local information on a fixed time scale. Recently, a multi-scale convolutional layer has been demonstrated to effectively exploit different scales of brain activity [35]–[37]. In this study, multi-scale 1D time-domain convolutional layers are designed for feature extraction. The output features can be expressed as follows:

\[
z_i = \text{Relu} (\hat{x}_i \cdot w)
\]

(3)

where \( \hat{x}_i \) is the input of the 1D convolutional layer of size \( 16 \times 400 \) and \( w \) denotes the convolutional kernel. \( \text{Relu}() \) is an activation function.

In our model, the features are extracted using three different temporal scales of kernels. The kernel sizes are 16 (channels) \( \times \) 4 (time points) \( \times \) 32 (filters), 16 \( \times \) 8 \( \times \) 32, and 16 \( \times \) 16 \( \times \) 32. It is of note that the padding operation is added to ensure the same length output of the convolutional layers. Furthermore, a concatenation layer is applied to merging the output features from convolutional layers. Subsequently, a max-pooling operation with a kernel size of \( 1 \times 5 \) is performed for dimensionality reduction. Finally, feature maps with a size of 96 \( \times \) 80 are fed into three fully connected layers. Finally, we use the softmax operation to estimate the classification probability.

D. Domain Adaptation Module

The model may not generalize well across subjects because of the individual differences of EEG feature distributions [13], [24], [25]. However, domain adaptation can eliminate the overall distribution discrepancy across domains, and reduce the classification error in the target domain. Here, we use the conditional domain adversarial network (CDAN) [31], which can model the feature representation \( f = F(x) \) and classifier prediction \( g = G(x) \) simultaneously. Loss function is composed of the single-trial classification error \( L_k \) and the discrepancy in feature distributions \( L_d \) as follows:

\[
L_k(G) = -\frac{1}{n_t} \sum_{i=1}^{n_t} v_i^* \cdot \log \left( G(x_i^*) \right)
\]

(4)

\[
L_d(D, G) = -\frac{1}{n_s} \sum_{i=1}^{n_s} \log \left[ D \left( f_s^i, g_s^i \right) \right]
- \frac{1}{n_t} \sum_{i=1}^{n_t} \log \left[ 1 - D \left( f_t^i, g_t^i \right) \right]
\]

(5)

where \( L_k \) encourages the task classification of the source label predictor using cross-entropy loss, and \( L_d \) helps the domain predictor to correctly distinguish the samples in the source domain from those in the target domain. In practice, the minimax problem of CDAN is as follows:

\[
\min_{G} L_k(G) - \lambda L_d(D, G)
\]

(6)

\[
\min_{D} L_d(D, G)
\]

(7)

where \( \lambda \) is a hyperparameter used to tradeoff single-trial classification and domain adversary. The joint distributions of \( f \) and \( g \) are modeled by conditioning domain predictor \( D \) on label prediction \( g \) with the multilinear map:

\[
D(f, g) = D(f \otimes g)
\]

(8)

In total, the minimax problem of CDAN can be formulated as:

\[
\min_{G} \frac{1}{n_s} \sum_{i=1}^{n_s} v_i^* \cdot \log \left( G(x_i^*) \right)
+ \lambda \left( \frac{1}{n_s} \sum_{i=1}^{n_s} \log \left[ D \left( f_s^i, g_s^i \right) \right] \right)
+ \lambda \left( \frac{1}{n_t} \sum_{i=1}^{n_t} \log \left[ 1 - D \left( f_t^i, g_t^i \right) \right] \right)
\]

(9)

\[
\min_{D} \frac{1}{n_s} \sum_{i=1}^{n_s} \log \left[ D \left( f_s^i, g_s^i \right) \right]
+ \frac{1}{n_t} \sum_{i=1}^{n_t} \log \left[ 1 - D \left( f_t^i, g_t^i \right) \right]
\]

(10)

E. Training Setup

In this experiment, we use the leave-one-subject-out method to evaluate the generalization performance of the model. Thus, the EEG samples from 16 subjects are divided into 16 folds. It should be noted that the trials from the same subject are not split across folds. Since there are only a few target samples for a single subject, we use random sampling and an ensemble learning approach to solve the data imbalance and improve classification robustness. During the training procedure, we randomly select non-target samples in the training set to ensure that the number of non-target samples is the same as that of the target samples. Subsequently, the trained model is used for the test. We repeat the random sampling 10 times. Notably, the sampling operation is not implemented in the test set. Therefore, we obtain the classification results of the 10 different models. The mean classification confidence
of these models is used as the final classification results. The same approach is replicated for all algorithms in this study.

Additionally, we choose the Adam optimizer to optimize our proposed model and cross-entropy as the loss function of source classifier. The batch size is empirically set as 16. Moreover, we use a dropout mechanism (dropout = 0.5) to avoid overfitting. The proposed model is coded based on Pytorch and trained on NVIDIA RTX 2080Ti GPU.

IV. EXPERIMENT

A. Subjects

Sixteen healthy subjects (eight professionals, age range: 19–35 years), with normal or corrected-to-normal vision, participated in the experiments. None of the subjects had a history of mental illness. Before the experiments, all participants were instructed to read the details of the experimental procedure and sign an informed consent.

B. Experiment Procedure

The AVEP paradigm uses remote sensing images (700 × 700 pixels) as the stimuli. Images with and without an airplane are considered as target stimuli and non-target stimuli, respectively. Each target image only contains only one target airplane. As shown in Fig. 2A, each subject should perform four sessions of the AVEP task and each session is composed of four blocks. During one block, each of 20 images (18 non-target images and 2 target images) is randomly presented at the center of the screen for 3 s at 0.5 s interval. Notably, the interval between the two target images in one block is at least 10 s. The subjects have a rest for 1 min between blocks and 5 min between sessions.

C. EEG Acquisition and Preprocessing

The EEG data is recorded using the standard amplifier Brain Amp (Brain Products, Germany) at a sampling frequency of 1000 Hz based on the BCI2000 system. The 16 electrodes are placed following the international 10-20 system (Fig. 2B). The EEG data are referenced to the right earlobe. The raw EEG data are then preprocessed using a simple preprocessing pipeline. First, EEG data are down-sampled to 125 Hz. Second, a FIR filter with a bandwidth of 0.1–40 Hz is applied to the EEG signals using EEGLAB [38]. Then, the EEG trials are extracted according to the onset of each image stimulus, and corrected using the average amplitude of the 200 ms preceding the stimulus onset as the baseline. Consequently, we obtain preprocessed EEG samples $D = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^{C \times L}$ denotes a data array of 16 electrodes × 400 sampling points, $y_i$ is the task label, and $n$ denotes the number of samples.

D. Comparison Models

We use typical EEG feature extraction based on SVM as a baseline method. To better exploit the brain responses in long-term EEG signals and improve model performance, we extract various handcrafted features as the input of the linear classifier. First, the EEG samples are divided into temporal segments using a sliding window approach with a step of one sampling point. The window size $W$ is set to 800 ms (100 sampling points) to include sufficient effective information of the ERP components [6]. We then obtain a series of temporal segments $s_j$, $j = 1, 2, ..., L − W + 1$. For each temporal segment $s \in \mathbb{R}^{C \times W}$, we calculate multiple handcrafted features for each channel:

1. Variance: $\sigma^2 = \frac{\sum_{k=1}^{W}(s(k) − \bar{s})^2}{100}$, where $k$ denotes the $k$th sampling point and $\bar{s}$ denotes the mean value of temporal segment $s$;
2. Amplitude range: $A = \max(s) − \min(s)$;
3. Energy: $E = \sum_{k=1}^{W}(s(k))^2$;
4. Frequency corresponding to maximum power spectral density: $PSE = −\sum_{i=1}^{m} p_i \ln p_i$, where $p_i (i = 1, 2, ..., m)$ is the relative power in the frequency range;
5. Singular spectrum entropy (SSE): the trajectory matrix is constructed from the temporal segments, $Y = [s_1, s_2, \cdots, s_{L−W+1}]^T$. Then, singular value decomposition can be expressed as: $Y = U SV^T$, where $U$ and $V$ are unitary matrices, and $S = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_R)$ is the diagonal matrix of $R$ nonzero singular values. The singular spectrum entropy can be written as $\text{SSE} = −\sum_{i=1}^{R} c_i \ln c_i$, where $c_i = \sigma_i / \sum_{i=1}^{R} \sigma_i$. Finally, features from different temporal segments are concatenated into a feature vector as the input of a linear SVM classifier.

In addition, the performance of the proposed model is compared with that of existing CNN-based approaches, such as EEGNet, ShallowConvNet, and DeepConvNet. EEGNet was proposed by Lawhern et al. [17] for BCI classification tasks and consists of three blocks. The first block has two convolutional steps: 2D convolution and depthwise convolution.
The second block is composed of a separable convolution (depthwise convolution and pointwise convolutions) and an average pooling layer. The last block is the classification block, which is a softmax classification layer. Here, we use two models (EEGNet-4,2 and EEGNet-8,2) with different numbers of filters, as described in a previous study [17]. EEGNet-4,2 represents four 2D convolutional filters and $4 \times 2$ depthwise convolutional filters in block 1, and EEGNet-8,2 represents eight 2D convolutional filters and $8 \times 2$ depthwise convolutional filters in block 1.

DeepConvNet and ShallowConvNet were designed by Schirrmeister et al. [39]. The DeepConvNet model is a generic architecture that includes five convolutional layers and a dense softmax classification layer. The ShallowConvNet architecture comprises two parts. The first part of the model consists of a temporal convolution and a spatial filter. The next part includes three components: a squaring nonlinearity, a mean pooling layer, and a logarithmic nonlinearity. Full details of DeepConvNet and ShallowConvNet can be found in the original paper [39]. These three models were also used in previous studies for 128 Hz EEG signals. Since the sampling rate of our data (125 Hz) is almost equal to 128 Hz, we use these models with the same hyperparameters as that of the previous study [17]. The same training approach is replicated for all the algorithms in this study.

### V. Results

#### A. Classification Performance

We compare the classification performance of our model with those of the conventional supervised classification method (SVM) and CNN-based algorithms (EEGNet, DeepConvNet and ShallowConvNet). The 16-fold cross-validation results of the cross-subject classification across all algorithms are shown in Table I. We compute the following metrics to measure the models’ performance: accuracy, F1-Score, true positive rate (TPR), false positive rate (FPR), and area under the curve (AUC).

Despite the use of multiple time-frequency analysis approaches for feature extraction, it can be seen that all the deep learning models show significant better performances than the baseline method. The results indicate that the SVM classification model has poor generalizability in cross-subject classification tasks because the classification hyperplane based on handcrafted features is more specific than the deep learning models. Moreover, compared with other CNN-based algorithms, our proposed DC-tCNN model shows a significant improvement in performance. We can see that our model achieves the best performance for the F1-score, TPR, and AUC. Specifically, the improvement is 6.18%–28.84% for F1-score and 5.57%–53.45% for TPR. The results show that ShallowConvNet outperforms EEGNet-4,2, whereas EEGNet-4,2 outperforms EEGNet-8,2. DeepConvNet performs the worst among these deep models. This may be because DeepConvNet with 5 convolution layers is more complicated than EEGNet and ShallowConvNet, demonstrating that deeper features are more task-specific than shallow features. The findings of previous works also suggest that shallower features learned from deep model may have better generalizability across tasks and subjects than deeper features [40], [41]. It should be noted that DeepConvNet achieves the best performance with an accuracy of 92.19% and FPR of 1.88%, but the TPR and F1-score are significantly lower than those of DC-tCNN, EEGNet, and ShallowConvNet. The results indicate that DeepConvNet overfits the non-target samples. The high performances of accuracy and FPR for DeepConvNet are caused by the imbalanced data in the test set.

#### B. Spatiotemporal Feature Analysis

To visualize the contribution of channel-wise attention, we calculate the average channel weights of the EEG trials for each subject. As illustrated in Fig. 3, the spatial topographies of the average channel weights are used to demonstrate the spatial features learned by our model. Note that the channel weights for different subjects are different, demonstrating that there is a discrepancy in the EEG feature distributions between the participants. The 16 electrodes are
placed following the international 10-20 system. For most subjects, it can be seen that the weights of parietal, central, and frontal regions are greater, which is consistent with previous studies [12], [16], [18], [42].

The EEG topographic maps of the target and non-target trials from one randomly selected subject are shown in Fig. 4. It can be found that the recorded brain activity is relatively stable during the period corresponding to the non-target trials (Fig. 4A). However, there is obvious brain activation over the parietal lobe in the target trials (Fig. 4B), indicating that the ERP components are evoked by the target stimulus. We can also see that the peak latencies of the different target trials are inconsistent. This observation suggests that the ERP responses induced by the dim target vary across trials, because the discovery time of the dim targets depends mainly on the participants.

Moreover, we visualize the hidden features of three different testing EEG samples from one cross-subject fold of the dataset and calculate the relevance of the features on the resulting classification decision. In this study, we use Captum [43] to compute the single-trial EEG feature relevance. Captum is a model interpretability library designed for Pytorch models, which can be used to visualize the feature contributions to the output predictions. Fig. 5A shows that the time points when features are activated are significantly different across trials, indicating that event-related brain responses lag behind stimulus onset time in a subject-driven manner. It is noteworthy that Captum feature relevance aligns closely with the activated hidden features (Fig. 5B), indicating that the model can exactly extract relevant and significant EEG features that accurately reflect event-related brain responses.

C. Ablation Analysis

We conduct an ablation analysis to investigate the necessity of using domain adaptation and channel attention modules. The classification results of ablation analysis are presented in Table II. We can see that tCNN without domain adaptation and channel attention modules also achieves considerable classification performance, with accuracy and AUC reaching 88.08% and 97.67%, respectively. The improvement is significant compared with SVM and other typical deep models, indicating the superior performance of the multi-scale convolutional layer in solving the unexplicit target onset time and inter-individual variability of ERPs. After applying the channel-wise attention mechanism to focus on the relatively important channels and reduce the interference of noise, C-tCNN obtains a subtle performance improvement compared with tCNN in terms of accuracy and FPR. Although both C-tCNN and tCNN obtain high classification accuracies, their TPR and F1-score performances are poor, indicating that these models cannot
predict the target samples accurately but overfit the non-target samples.

Furthermore, compared with C-tCNN, D-tCNN achieves a significant performance improvement from 62.47% to 65.68% for the F1-Score and 79.6% to 93.9% for the TPR. This indicates that the inclusion of the domain adaptation module can reduce the discrepancy in the feature distributions, especially for target samples. Additionally, after jointly adding channel-wise attention and domain adaptation modules, the results indicate that DC-tCNN obtains the highest accuracy and F1-Score. This suggests that the proposed model can have a better generalizability and classification performance for the cross-subject classification task.

To visualize the feature distributions of the source and target domains in these four models, t-distributed Stochastic Neighbor Embedding (t-SNE) is applied to project the embedded features into two dimensions. As illustrated in Fig. 6, the feature distributions of the source and target domains in DC-tCNN and D-tCNN are more consistent than those in C-tCNN and tCNN. The distribution discrepancy is reduced after adding the domain adaptation module, which further demonstrates the effectiveness of domain adaptation in eliminating the distribution discrepancy.

VI. DISCUSSION

In this study, we develop a novel BCI paradigm called AVEP to detect dim targets in remote sensing images. The proposed AVEP paradigm has three main advantages. First, the proposed paradigm might be useful for detecting dim targets in large high-resolution images by making full use of the advantages of human visual perception. The RSVP paradigm is more suitable for detecting conspicuous targets, such as experiments of finding target images containing people [8], [13], [14]. Although Bigdely-Shamlo et al. [11] designed an RSVP experiment to find target airplanes in satellite images, the performance and efficiency of detection are highly dependent on preprocessing steps, such as segmentation. Second, it is difficult for current computer vision algorithms to deal with dim target detection tasks, especially for new targets and complex backgrounds. In contrast, the proposed BCI-based systems are relatively reliable because brain responses are specific only to target attendance. Target images can be accurately identified by exploring their corresponding brain responses. Third, although RSVP uses a relatively high presentation rate, expert saccade-based visual search is demonstrated to be faster than RSVP-based search [11]. In particular, the areas that are unlikely to contain targets can be easily excluded by human experts. Therefore, the proposed paradigm may be superior to the RSVP in dim target detection tasks.

A longer search period for image targets is required because of the varying orientation, small size of the targets, and complicated background. Target onset time is asynchronous with stimulus onset time and varies across trials and subjects due to participant-dependent cognitive processes in the target search. Thus, it is challenging to detect specific ERP responses for dim target detection using averaged multiple ERP responses because of the absence of target onset time. To address this challenge, Song et al. designed a detection framework for video target detection by aligning EEG signals on a common template [44]. In contrast, we propose an end-to-end deep model DC-tCNN to solve this problem by combining multi-scale convolution, channel-wise attention, and domain adaptation modules. The experiments demonstrate that our proposed model can automatically explore underlying brain responses without alignment based on a single trial.

We also compare our model with conventional methods and state-of-the-art deep learning models for single-trial EEG classification. The results indicate that deeper models are more likely to exhibit lower performance in the cross-subject classification task. This is because generalizability decreases as the model becomes deeper. It is noteworthy that the classification performance of EEGNet-4,2 is better than that of EEGNet-8,2. One possible explanation is the number of model parameters. More learnable parameters of EEGNet-8,2 are more likely to overfit the limited training data, which is consistent with
the findings of previous work [17]. Our model achieves the best performance on single-trial EEG classification for the AVEP task, which indicates that the proposed model can effectively solve the unexplicit target onset time and cross-subject distribution discrepancy. In addition, to validate whether the proposed model learns neurophysiological features or only noise and artifact signals, we visualize the hidden features and calculate the single-trial feature relevance for one test subject. The results show that the features extracted by the DC-tCNN model are interpretable features rather than noise or artifact signals. Ablation analysis also demonstrates the necessity of each module in the model. Although adversarial domain adaptation is similar to GAN networks by integrating adversarial learning in the models, vanishing gradients may hardly occur because the loss function of domain adaptation includes not only discrepancies in feature distributions but also classification errors. Furthermore, in this study, we apply a novel conditional domain predictor to improve discriminability by capturing the cross-covariance dependency between the class predictions $g$ and feature representations $f$. The visualization results of the feature distributions of the source and target domains demonstrate the effectiveness of domain adaptation in eliminating distribution discrepancy.

In general, our work can provide a valuable reference for the BCI-based image detection of dim targets. Moreover, the AVEP paradigm and DC-tCNN model can be applied to other EEG-based BCI systems, such as auditory target detection. Importantly, we indicate the merit of this deep learning model in terms of generalizability, i.e., the trained model can be effectively applied to unseen subjects, even considering individual differences in cognitive processes. However, some limitations should be considered. The search for relatively small targets in large high-resolution images may lead to mental or visual fatigue. Therefore, the search process would be slower than ideal conditions because of frequent rest breaks. In realistic applications, for example, the locations of targets must be marked in the images. BCI recording combined with an eye tracker will address this issue in future work.

VII. Conclusion

In this study, we propose a novel BCI paradigm called AVEP for dim target detection in remote sensing images. A deep learning model of DC-tCNN is developed to achieve high-accuracy classification of the single-trial EEG data of the AVEP task. The experimental results indicate that DC-tCNN has superior performance and better generalizability than conventional linear methods and typical EEG deep learning networks. Visualization analyses of spatiotemporal features also demonstrate the interpretability and effectiveness of our proposed paradigm and algorithm for EEG-based dim target detection tasks. The AVEP paradigm and DC-tCNN model can effectively explore unexplicit event-related brain responses and may be used in other BCI systems such as auditory target detection. Overall, our work can provide a valuable reference and a step towards practical implementation for BCI-based image detection of dim targets.
[24] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, “Domain adaptation techniques for EEG-based emotion recognition: A comparative study on two public datasets,” *IEEE Trans. Cogn. Devel. Syst.*, vol. 11, no. 1, pp. 85–94, Mar. 2019.

[25] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, “Domain adaptation for EEG emotion recognition based on latent representation similarity,” *IEEE Trans. Cogn. Devel. Syst.*, vol. 12, no. 2, pp. 344–353, Jun. 2020.

[26] Y. Ming et al., “Subject adaptation network for EEG data analysis,” *Appl. Soft Comput.*, vol. 84, Nov. 2019, Art. no. 105689.

[27] Y. Qi, L. Ding, Y. Wang, and G. Pan, “Learning robust features from nonstationary brain signals by multi-scale domain adaptation networks for seizure prediction,” *IEEE Trans. Cognit. Develop. Syst.*, early access, Jul. 26, 2021, doi: 10.1109/TCDS.2021.3100270.

[28] J. Su, H. Shen, L. Peng, and D. Hu, “Few-shot domain-adaptive anomaly detection for cross-site brain images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 8, 2021, doi: 10.1109/TPAMI.2021.3125686.

[29] M. Long, Y. Cao, J. Wang, and M. Jordan, “Learning transferable features with deep adaptation networks,” in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[30] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer feature learning with joint distribution adaptation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2200–2207.

[31] M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Conditional adversarial domain adaptation,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1640–1650.

[32] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” 2014, arXiv:1411.1784.

[33] I. Daly, A. Matran-Fernandez, D. Valeriani, M. Lebedev, and A. Kübler, *Datasets for Brain–Computer Interface Applications*. Lausanne, Switzerland: Frontiers Media, 2021.

[34] S. Sur and V. Sinha, “Event-related potential: An overview,” *Ind. Psychiatry J.*, vol. 18, no. 1, p. 70, 2009.

[35] L. Fan, J. Su, J. Qin, D. Hu, and H. Shen, “A deep network model on dynamic functional connectivity with applications to gender classification and intelligence prediction,” *Frontiers Neurosci.*, vol. 14, p. 881, Aug. 2020.

[36] S. Roy, I. Kiral-Kornek, and S. Harrer, “ChronoNet: A deep recurrent neural network for abnormal EEG identification,” in *Artificial Intelligence in Medicine*. Berlin, Germany: Springer, 2019, pp. 47–56.

[37] W. Yan et al., “Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site FMRI data,” *EBioMedicine*, vol. 47, pp. 543–552, 2019.

[38] A. Delorme and S. Makeig, “EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis,” *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[39] R. T. Schirrmeister et al., “Deep learning with convolutional neural networks for EEG decoding and visualization,” *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[40] O. Mayer, B. Bayar, and M. C. Stamm, “Learning unified deep-features for multiple forensic tasks,” in *Proc. 6th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2018, pp. 79–84.

[41] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3320–3328.

[42] J. Shamwell, H. Lee, H. Kwon, A. R. Marathe, V. Lawhern, and W. Nothwang, “Single-trial EEG RSVP classification using convolutional neural networks,” *Proc. SPIE*, vol. 9836, pp. 373–382, May 2016.

[43] N. Kokhlikyan et al., “Captum: A unified and generic model interpretability library for PyTorch,” 2020, arXiv:2009.07896.

[44] X. Song, B. Yan, L. Tong, J. Shu, and Y. Zeng, “Asynchronous video target detection based on single-trial EEG signals,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 9, pp. 1931–1943, Sep. 2020.