A³CLNN: Spatial, Spectral and Multiscale Attention ConvLSTM Neural Network for Multisource Remote Sensing Data Classification

© 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

DOI:10.1109/TNNLS.2020.3028945

Heng-Chao Li, Senior Member, IEEE, Wen-Shuai Hu, Wei Li, Senior Member, IEEE, Jun Li, Senior Member, IEEE, Qian Du, Fellow, IEEE, and Antonio Plaza, Fellow, IEEE

A³CLNN: Spatial, Spectral and Multiscale Attention ConvLSTM Neural Network for Multisource Remote Sensing Data Classification

Heng-Chao Li, Senior Member, IEEE, Wen-Shuai Hu, Wei Li, Senior Member, IEEE, Jun Li, Senior Member, IEEE, Qian Du, Fellow, IEEE, and Antonio Plaza, Fellow, IEEE

Abstract—The problem of effectively exploiting the information multiple data sources has become a relevant but challenging research topic in remote sensing. In this paper, we propose a new approach to exploit the complementarity of two data sources: hyperspectral images (HSIs) and light detection and ranging (LiDAR) data. Specifically, we develop a new dual-channel spatial, spectral and multiscale attention convolutional long short-term memory neural network (called dual-channel A^{3} CLNN) for feature extraction and classification of multisource remote sensing data. Spatial, spectral and multiscale attention mechanisms are first designed for HSI and LiDAR data in order to learn spectral- and spatial-enhanced feature representations, and to represent multiscale information for different classes. In the designed fusion network, a novel composite attention learning mechanism (combined with a three-level fusion strategy) is used to fully integrate the features in these two data sources. Finally, inspired by the idea of transfer learning, a novel stepwise training strategy is designed to yield a final classification result. Our experimental results, conducted on several multisource remote sensing data sets, demonstrate that the newly proposed dual-channel A^{3} CLNN exhibits better feature representation ability (leading to more competitive classification performance) than other state-of-the-art methods.

Manuscript received February 7, 2020; revised August 25, 2020; accepted September 24, 2020. Date of publication October 21, 2020; date of current version February 4, 2022. This work was supported by the National Natural Science Foundation of China under Grant 61871335, in part by the Fundamental Research Funds for the Central Universities under Grant 2682020XG02 and 2682020ZT35. (Corresponding author: Wen-Shuai Hu.)

Heng-Chao Li and Wen-Shuai Hu are with the Sichuan Provincial Key Laboratory of Information Coding and Transmission, Southwest Jiaotong University, Chengdu 611756 China (e-mail: lihengchao_78@163.com; wshu@my.swjtu.edu.cn).

Wei Li is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081 China.

Jun Li is with the Guangdong Provincial Key Laboratory of Urbanization and Geo-simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275 China.

Qian Du is with the Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS 39762 USA.

Antonio Plaza is with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, 10060 Cáceres, Spain. *Index Terms*—Multisource remote sensing data classification, convolutional long short-term memory, attention mechanism, transfer learning, feature extraction, fusion.

I. INTRODUCTION

WITH the development of remote sensing technology, different sources of complementary data are now available from a variety of sensors. Hyperspectral images (HSIs) provide plenty of spectral information and have been widely used for land-cover classification purposes [1]-[2]. Different from HSI data, light detection and ranging (LiDAR) data consist of detailed elevation information. These data convey rich information in the spatial domain that can be used to improve the characterization of HSI scenes [3]-[4], as the LiDAR data are less affected by atmospheric interferers [5]. In the literature, several works [6]-[9] have discussed the fusion of multisource remote sensing data.

Many classification methods have been proposed to exploit the spatial-spectral information contained in HSI data, including machine learning-based methods [10], [11], tensor-based algorithms [12], sparse representationbased methods [13]. In recent years, deep learningbased algorithms have achieved great success in remote sensing data interpretation. Convolutional neural networks (CNNs) were first adopted for HSI classification by Hu et al. [14] and Chen et al. [15]. After these seminal works, many other deep learning-based HSI classification methods have been proposed, and these methods have been shown to be able to provide higher classification accuracies than traditional methods. Relevant examples are the CNN-based pixel-pair model [16], a spatial-spectral feature based classification (SSFC) model that stacks CNNs with balanced local discriminant embedding [17], or the siamese CNN-based method [18]. In addition to the CNN-based HSI classification algorithms, recurrent neural networks (RNNs) [19] have also achieved great success in the task of capturing useful information from different kinds of inputs, due to their unique ability for modeling longrange dependencies. As a result, many HSI classification models have been developed, including the RNNbased pixel-level spectral classification model [20], a local spatial sequential RNN (LSS-RNN) model [21], and a classification model that combines CNNs and RNNs [22]. Furthermore, in order to solve the gradient vanishing or explosion problem in RNNs, long shortterm memories (LSTMs) [23] and convolutional LSTMs (ConvLSTMs) [24] were proposed. The most common way to utilize them is in combination with a CNN, such as the convolutional RNN (CRNN) model for spectralcontextual feature extraction [25], a recurrent threedimensional (3-D) fully convolutional network [26], and a two-stage classification model that combines a 3-D CNN and and a (2-D) ConvLSTM [27]. In addition, there are also some relevant works focused on building deep feature extraction and classification models using the LSTM and (2-D) ConvLSTM cells as basic units, such as a spatial-spectral LSTMs (SSLSTMs) [28], the bidirectional-ConvLSTM (Bi-CLSTM) [29], or a spatialspectral ConvLSTM 2-D neural network (SSCL2DNN) [30]. These methods have achieved good performance in the task of HSI data classification. In order to better preserve the intrinsic structure of HSI data, a 3-D ConvLSTM cell was developed from the (2-D) ConvLSTM cell in [30], from which a spatial-spectral ConvLSTM 3-D neural network (SSCL3DNN) was designed.

Considering the special characteristics of HSI and LiDAR data, several works aimed at fusing these two data sources in order to improve the classification performance [9]. Examples include decision-fusion classification methods [6], [8] and morphological feature extraction-based algorithms [31]-[33]. In addition, several deep learning-based classification methods have also been proposed for the classification of multisource remote sensing data. Xu et al. [34] built a two-branch CNN model with data augmentation for fusing HSI and LiDAR features. In [35], an unsupervised patch-to-patch CNN (PToP CNN) model was designed for HSI and LiDAR data classification, in which a three-tower PToP mapping is used to fuse their multiscale features. By introducing maximum correntropy criterion, Li et al. [36] proposed a dual-channel robust capsule network (dualchannel CapsNet) for the fusion of HSI and LiDAR data. However, it should be noted that a fixed size of the convolution kernel is used for all classes, which may lead to the absence of multiscale information for different classes. In this case, the complementarity of HSI and LiDAR data is not fully exploited, since spectral and spatial information cannot be effectively integrated.

The attention mechanism is an important technique derived from computational neuroscience [37]. It allows a given model to automatically locate and capture the significant information from the input. Since Bahdanau

et al. [38] firstly utilized it to select reference words from source sentences, numerous works have demonstrated that, with the help of attention mechanisms, deep learning-based models can obtain better feature representation ability in many fields of computer vision [39]-[45]. Several works have applied attention mechanisms to remote sensing problems. Cui et al. [46] proposed a dense attention pyramid network for ship detection in synthetic aperture radar (SAR) images, in which a convolutional block attention module (with spatial and channel-wise attention) is designed for highlighting salient features at specific scales. Chen et al. [47] improved the faster region-based CNN by using multiscale (spatial) and channel-wise attention for object detection in remote sensing images. With a skip-connected encoder-decoder model, the work in [48] developed an end-to-end multiscale visual attention network for highlighting objects and suppressing background regions. Wang et al. [49] proposed an end-to-end attention recurrent CNN for classification of very high-resolution (VHR) remote sensing scenes. Regarding HSI classification tasks, an attention-based inception model was designed in [50] which can accurately model spatial information in HSI data. Mou et al. [51] put forward a learnable spectral attention module (prior to CNN-based classification) for selecting informative bands. By combining an attention mechanism and RNNs, a spatial-spectral visual attentiondriven feature extraction model was designed in [52].

In this paper, a new dual-channel spatial, spectral and multiscale attention ConvLSTM neural network (dual-channel A^{3} CLNN) model is developed for the classification of multisource (i.e., HSI and LiDAR) remote sensing data. Specifically, three types of attention mechanisms are designed for extracting spectraland spatial-enhanced multiscale features. Furthermore, a novel three-level fusion strategy is designed for effectively integrating the information coming from the HSI and LiDAR data. In the first-level fusion stage, composite attention learning is proposed for fully exploiting spatial and spectral information in the LiDAR and HSI data. Then, both types of features are cascaded as the input of the classification layer, which is the intermediate stage. Since the order of magnitude of HSI features is much larger than that of the LiDAR features, in the third-level fusion stage the LiDAR features are reused at the top of a fusion network to make full use of the LiDAR data source on the classification performance. To effectively train the proposed model, a stepwise training strategy is designed, in which these two branches (HSI and LiDAR) are first trained individually to obtain the primary features, and then -inspired by the idea of transfer learning [53]- these features are used for initializing a fusion network that extracts high-level features. Finally, a multi-task loss function is designed to achieve a better optimization of the proposed dualchannel A^{3} CLNN model. The main contributions of this work can be summarized as follows.

- (1) Considering the wealth of spectral and spatial information presented in HSI and LiDAR data, we develop novel and learnable spectral and spatial attention modules to obtain spectral- and spatialenhanced features.
- (2) For different classes, a fixed-scale feature extraction strategy may be inappropriate due to the different scale information contained in these classes. To solve this problem, a learnable multiscale residual attention module is further designed that enhances the multiscale information representation ability of the whole model.
- (3) A three-level fusion strategy is proposed. Particularly, a composite attention learning module that combines spectral and spatial attention is designed as a two-level attention strategy that makes better use of the spectral and spatial information. In the training stage, a stepwise training strategy (with a multi-task loss function) is designed for optimizing the proposed dual-channel A³CLNN model, which can effectively accelerate its convergence speed.

The remainder of the paper is organized as follows. Section II reviews the ConvLSTM2D, ConvLSTM3D, and the attention mechanism. In Section III, the proposed dual-channel A^{3} CLNN model is described in detail. An exhaustive analysis of parameter settings and a quantitative evaluation of the proposed model are given in Section IV. Section V concludes the paper with some remarks and hints at plausible future research lines.

II. RELATED WORK

A. ConvLSTM2D and ConvLSTM3D

As a modification and an extended version of LSTM, Shi *et al.* [24] developed a ConvLSTM cell by extending the data processing method in LSTM to a 2-D convolution operation, with which plenty of the ConvLSTMbased deep models have been built for HSI classification [25], [26], [27], [29], [30]. Inspired by [30], it is further named ConvLSTM2D cell for convenience. However, as shown in [29] and [30], due to the special structure of the ConvLSTM2D cell, 3-D HSI data must be decomposed into a 2-D sequence when used as the input of the model, which may lose the intrinsic structure of HSI data.

To better preserve the intrinsic structure of HSI data, the ConvLSTM3D cell is further extended from the ConvLSTM2D cell, with which Hu *et al.* [30] proposed a novel and effective SSCL3DNN model for HSI classification. Nevertheless, there is still much room to further improve the performance of SSCL3DNN. For example, the multiscale information is not considered, and the characteristics of the ConvLSTM3D layer for modeling long-term dependencies are not fully utilized.

B. Attention Mechanism

After the attention mechanism was first introduced into deep learning in [38], an increasing number of attention-driven deep learning-based models have been proposed. These models were able to improve the feature representation ability in many research fields. In [39], Vaswani *et al.* introduced the following equation to calculate the output of the attention mechanism:

$$Attention(Q, K, V) = softmax(f(Q, K))V, (1)$$

where $f(\cdot)$ is the attention function, and Q, K, and V are the inputs. $softmax(\cdot)$ denotes the softmax function used for normalization. Specifically, there are two common attention functions, i.e., additive attention [38] and dot-product attention [39], where their corresponding definitions can be written as:

$$f_{additive}(Q, K) = W_Q Q + W_K K$$
$$f_{dot-product}(Q, K) = Q K^T,$$
(2)

where W_Q and W_K are the parameter weights, and T represents the transpose of the matrix.

The most common way for introducing an attention mechanism into deep learning is to build a hard part selection subnetwork or a soft mask branch [41]. By using residual learning, a residual attention module was built for soft pixel-level attention learning in [42], and then applied to image classification. In addition, channelwise attention [43], spatial and temporal-wise attention [44], spatial and channel-wise attention [46], and spatial-spectral attention [52] have also been proposed for feature enhancement. However, to the best of our knowledge, there have been no effective implementations of an attention mechanism for fusion and classification of multisource remote sensing data. In the following section, we develop a new composite attention learning module that combines spatial and spectral attention learning and a multiscale (residual) attention learning module for effectively combining HSI and LiDAR data.

III. DUAL-CHANNEL A^3 CLNN

A. Architecture Overview

It is well known that HSI data consist of many bands carrying plenty of spectral information, while LiDAR data are rich in height (spatial) information [34]. This motivates us to build a classification model upon a twobranch framework that fully exploits the complementary information from both sources of information.

The overall framework of the proposed dual-channel A^3 CLNN model is graphically represented in Fig. 1. A spectral attention block (SeAB) and a spatial attention block (SaAB) are first proposed for composite attention learning (see subsection III-B). A multiscale residual attention block (MSRAB) is designed in subsection III-C. In subsection III-D and III-E, the HSI branch (marked



Fig. 1. Architecture of the proposed dual-channel A^{3} CLNN model.



Fig. 2. Structure of the spectral attention block (SeAB).



Fig. 3. Structure of the spatial attention block (SaAB).



Fig. 4. Proposed multiscale residual attention block (MSRAB).

with yellow arrows in Fig. 1) and the LiDAR branch (marked with blue arrows in Fig. 1) are described in detail. The proposed three-level fusion strategy (marked with red arrows in Fig. 1) is described in subsection III-F. Finally, subsection III-G describes the multi-task loss function and the stepwise training strategy.

B. Composite Attention Learning

1) Spectral Attention Block (SeAB): An effective and learnable SeAB module is designed to learn more discriminative and spectral-enhanced feature representation. The structure of this module is described in Fig. 2.

Let $X_l^H \in R^{w_l \times h_l \times s_l \times c_l}$ denote the output of the *l*th ConvLSTM3D layer (or the original HSI data), where s_l, w_l, h_l , and c_l are respectively the number of spectral bands, width, height, and channel number. The purpose of SeAB is to learn an attention vector α_{Se}^H .

As shown in Fig. 2, a 3×3 and a 1×1 ConvL-STM2D layers, a spatial pooling layer, and a softmax function comprise the main backbone of the proposed SeAB module. Firstly, X_l^H is decomposed into s_l 2-D components along the spectral dimension, and converted into a sequence, i.e., $\left\{X_{l1}^{H}, \ldots, X_{lt}^{H}, \ldots, X_{ls_l}^{H}\right\}, t \in$ $\{1, 2, \ldots, s_l\}$, which are then fed (one by one) to the 3×3 ConvLSTM2D layer to model the long-range dependencies in the spectral domain. Another 1×1 ConvLSTM2D layer is added to generate an unnormalized attention map Z_{Se}^{H} , with size $w_l \times h_l \times s_l \times 1$. A spatial pooling operation is then applied to Z_{Se}^{H} to transform it into an unnormalized attention vector z_{Se}^{H} , with length s_l . Finally, z_{Se}^H is fed to a softmax function to yield the normalized attention vector α_{Se}^{H} , which is multiplied by the input to yield a spectral-enhanced feature representation. The output X_{l}^{H} of SeAB can be described as:

$$z_{Se}^{H} = f_p(f_{CL2D1}(f_{CL2D3}(X_l^H)))$$

$$\alpha_{Se}^{H} = softmax(z_{Se}^H)$$

$$\hat{X_l^H} = X_l^H \odot \alpha_{Se}^H,$$
(3)

where \odot is an element-based product operation. $f_{CL2D3}(\cdot)$ and $f_{CL2D1}(\cdot)$ respectively indicate the 3×3 and 1×1 ConvLSTM2D layers. $f_p(\cdot)$ is a spatial pooling layer with size $w_l \times h_l$, and $softmax(\cdot)$ represents the softmax function. Particularly, the dimension time_step in each ConvLSTM2D layer is set to s_l .

It should be noted that the input of SeAB can be either the original data or the output of the last layer. The special structure of SeAB makes it a feature enhancement module that can be added to any layer of the whole network to obtain spectral-enhanced feature representation. In our experiments, SeAB is added to the ConvLSTM3D layer.

2) Spatial Attention Block (SaAB): Unlike HSI data, LiDAR data provide rich elevation information, which means information about the height and shape of targets [34]. We design an SaAB module for exploiting the spatial information in LiDAR data, which will lead to more effective spatial-enhanced feature representation.

Let $X_l^L \in R^{w_l \times h_l \times c_l}$ be the output of the *l*th ConvLSTM2D layer (or the original LiDAR data), in which w_l , h_l , and c_l are defined in accordance with those in the SeAB. SaAB is constructed to learn an attention map α_{Sa}^L .

The structure of the proposed SaAB module is shown in Fig. 3. The main backbone of this module is given by a 3×3 and a 1×1 ConvLSTM2D layer, and a softmax function. Firstly, the 3×3 and 1×1 ConvLSTM2D layers are used to generate an unnormalized attention map Z_{Sa}^L with size $w_l \times h_l \times 1$. After that, a softmax function is utilized to generate a normalized attention map α_{Sa}^L . The forward propagation of SaAB can be written as:

$$Z_{Sa}^{L} = f_{CL2D1}(f_{CL2D3}(X_{l}^{L}))$$

$$\alpha_{Sa}^{L} = softmax(Z_{Sa}^{L})$$

$$\hat{X_{l}}^{L} = X_{l}^{L} \odot \alpha_{Sa}^{L},$$
(4)

where X_l^L denotes the output of SaAB. The definitions of \odot , $f_{CL2D3}(\cdot)$, $f_{CL2D1}(\cdot)$, and $softmax(\cdot)$ are similar to that in (3). However, different from SeAB, the dimension time_step in the SaAB module is fixed to 1.

Similar to SeAB, SaAB can be treated as an effective spatial feature extractor, and used by any layer of a deep learning-based model to obtain spatial-enhanced features. In our experiment, SaAB is utilized after the ConvLSTM2D layer in the LiDAR branch.

Based on the two aforementioned attention blocks, an effective composite attention learning approach is proposed for jointly learning the spatial-spectral features. Our approach can efficiently exploit spectral and spatial information, and enhance the feature extraction ability of the whole model. A more detailed description of the attention mechanism is given in subsection III-F.

C. Multiscale Residual Attention Block (MSRAB)

In multisource remote sensing data, different classes may comprise different scale information, which means classification models using uniform scale to extract features may not meet the scaling requirements of different classes. Therefore, it is necessary to design a multiscale feature extractor able to properly describe multiscale information. An MSRAB module (integrating residual learning and attention mechanism) is proposed. Taking the LiDAR branch as an example. Let $X_l^L \in R^{\tau_l \times w_l \times h_l \times c_l}$ and X_{l+1}^L denote the input and output of the MSRAB module, where τ_l is the dimension time_step of the ConvLSTM2D layer. The structure of the MSRAB module is given in Fig. 4. In particular, the first column is the multiscale feature extraction function realized by the ConvLSTM2D layers, which uses different fields of perception to capture multiscale information from the input (using different scales such as 1×1 , 3×3 , and 5×5). These features are cascaded in the time_step dimension of the ConvLSTM2D layer. After the ConvLSTM2D layers, the obtained multiscale features are further learned in non-linear fashion, yielding an unnormalized attention map $Z_{MSR}^L \in R^{3\tau_l \times w_l \times h_l \times c_l}$. Then, after a global average pooling (GAP) layer, the unnormalized attention map Z_{MSR}^L is converted to an unnormalized attention vector z_{MSR}^L with length $3\tau_l$, which is further normalized by a softmax function to generate a multiscale attention vector α_{MSR}^L . The output O_{MSR}^L of the multiscale attention part in the MSRAB module is written as:

$$T = [f_1(X_l^L), f_3(X_l^L), f_5(X_l^L)]$$

$$z_{MSR}^L = f_p(f_{CL2D1}(f_{CL2D3}(T)))$$

$$\alpha_{MSR}^L = softmax(z_{MSR}^L)$$

$$O_{MSR}^L = T \odot \alpha_{MSR}^L, \qquad (5)$$

where $[\cdot, \cdot, \cdot]$ denotes the concatenation operation, and $f_1(\cdot)$, $f_3(\cdot)$, and $f_5(\cdot)$ are the multiscale feature extraction functions realized by the ConvLSTM2D layer.

Unlike the cascading approach in [54] and [55], the output of MSRAB is fed to a fusion layer (built by a ConvLSTM2D layer) to model long-term dependencies in the multiscale dimension. Furthermore, residual learning is also applied to mitigate the gradient vanishing or explosion problems through a feature reuse mechanism. The forward propagation of MSRAB is expressed as:

$$X_{l+1}^{L} = f_{CL2Da}(O_{MSR}^{L}) + X_{l}^{L},$$
 (6)

where $f_{CL2Da}(\cdot)$ denotes an $a \times a$ ConvLSTM2D layer.

Specially, the structure of MSRAB in the HSI branch is similar to that in the LiDAR branch. However, the dimensions X_l^H and X_{l+1}^H need to be extended to $R^{\tau_l \times w_l \times h_l \times s_l \times c_l}$ and $R^{3\tau_{l+1} \times w_{l+1} \times h_{l+1} \times s_{l+1} \times c_{l+1}}$, in which s_l and s_{l+1} are the spectral dimensions. $f_1(\cdot)$, $f_3(\cdot)$, $f_5(\cdot)$, $f_{CL2D1}(\cdot)$, $f_{CL2D3}(\cdot)$, and $f_{CL2Da}(\cdot)$ in (5) and (6) are implemented by ConvLSTM3D layers.

Similar to SeAB and SaAB, MSRAB can be used as a multiscale information enhancement module to bring a larger receptive field to the whole model, and MSRAB can adaptively focus on important areas at each scale.

D. Multiscale Spectral Attention Neural Network (MSSeA) for the HSI Branch

An MSSeA model is proposed for HSI branch, which is marked in Fig. 1 by yellow arrows. The backbone of MSSeA consists of a ConvLSTM3D layer, an SeAB module, a pooling layer, an MSRAB model, a GAP layer, and a classification layer. Specifically, considering the redundant information presented in the original HSI data, principal component analysis (PCA) is used for spectral dimensionality reduction in our experiments.

Let $W \times H \times D$ denote the size of the original HSI data, where W, H, and D are the width, height, and the number of the spectral bands, respectively. In the data preprocessing stage, the first K principal components are retained, and the pixels in a local neighborhood window with size $s \times s$ are extracted to account for the spatialcontextual information around each pixel x. Accordingly, the whole data associated to pixel x can be represented as $X^H \in R^{s \times s \times K}$, which is also the input of MSSeA. To make the data meet the format requirements of the ConvLSTM3D layer, X^H is decomposed into τ 3-D components and then converted into a sequence with length τ , as indicated below:

$$X^{H} \Rightarrow \left\{ X_{1}^{H}, \dots, X_{t}^{H}, \dots, X_{\tau}^{H} \right\}, \tag{7}$$

where X_t^H is the *t*th 3-D component, and $t \in \{1, 2, ..., \tau\}$. τ is the dimension time_step in the ConvLSTM3D layer (fixed here to 1).

Then, this sequence is fed into l cascaded ConvL-STM3D layers (one by one) to extract shallow spatialspectral features. To facilitate subsequent variable representation, the output of the *l*th ConvLSTM3D layer is written as $X_l^H \in R^{\tau_l \times w_l \times h_l \times s_l \times c_l}$, where τ_l is the dimension time_step, and the size of the convolution kernel is $k_l^H \times k_l^H \times k_l^H$. Following each ConvLSTM3D layer, SeAB is applied to extract the spectral-enhanced features and, according to (3), the enhanced features can be expressed as X_l^H . In our experiments, l is set to 1.

Furthermore, to measure the multiscale information and meet the scale requirements of different classes, the spectral-enhanced features X_l^H , obtained after a pooling layer, are input to an MSRAB module to learn the multiscale information. From (6), the extracted multiscale features can be written as $X_{l+1}^H \in R^{3\tau_{l+1} \times w_{l+1} \times h_{l+1} \times s_{l+1} \times c_{l+1}}$. Inspired by [56] and for the sake of accelerating convergence and solving the gradient vanishing problem, a batch normalization (BN) layer and a swish function are used for regularization.

Then, we apply a GAP layer [57] at the top of MSSeA –instead of the fully connected (FC) layer– to map the feature space to class label space, which can directly endow each channel with the actual category meaning, regularize the whole model, and prevent over-fitting to some degree. In addition, this strategy can effectively

solve the problem of having too many parameters in the FC layer, relaxing the limitations imposed to the model by the resolution of the input. The forward propagation of the GAP layer in MSSeA can be expressed as:

$$X_{GAP}^{H} = f_{GAP}(X_{l+1}^{H}),$$
 (8)

where $f_{GAP}(\cdot)$ and $X_{GAP}^{H} \in R^{1 \times c_{l+1}}$ denote the expression and output of the GAP layer, respectively.

Finally, a classification layer (with the softmax function) follows the GAP layer to predict the conditional probability distribution $P_c^H = P(y = c | X_{GAP}^H, W, b) = \frac{e^{(W_c X_{GAP}^H + b_c)}}{\sum_{j=1}^{N} e^{(W_j X_{GAP}^H + b_j)}}$ of each class c, where $c \in 1, 2, \ldots, N$, and N denotes the number of classes.

To obtain the final classification results, the cross entropy is selected as the loss function to optimize the HSI branch, which is named $Loss_H$ for convenience.

E. Multiscale Spatial Attention Neural Network (MSSaA) for the LiDAR Branch

Similar to subsection III-D, an MSSaA model is designed for the LiDAR branch (marked in Fig. 1 with blue arrows). A ConvLSTM2D layer, an SaAB module, a down-sample layer, an MSRAB module, a GAP layer, and a classification layer represent the backbone of it.

Let us assume that the size of the original LiDAR data is $W \times H$. In the data preparation stage, a $s \times s$ local spatial-contextual window around each pixel x is first extracted, which is fed into the MSSaA and expressed as $X^L \in R^{s \times s}$. Due to the special structure of the ConvLSTM2D layer, X^L needs to be decomposed into a sequence with τ 2-D components as follows:

$$X^{L} \Rightarrow \left\{ X_{1}^{L}, \dots, X_{t}^{L}, \dots, X_{\tau}^{L} \right\},$$
(9)

where X_t^L is the *t*th 2-D component and $t \in \{1, 2, ..., \tau\}$. τ is the dimension time_step in the ConvLSTM2D layer, which is set to 1 in our experiments.

Then, this sequence is fed into l cascaded ConvL-STM2D layers (one by one) to extract the shallow spatial features. For convenience, the output of the lth ConvLSTM2D layer is described as $X_l^L \in R^{\tau_l \times w_l \times h_l \times c_l}$. After that, an SaAB module is applied to enhance the spatial information of the output of each ConvL-STM2D layer. According to (4), the enhanced features are written as \hat{X}_l^L . Furthermore, a pooling layer and an MSRAB are utilized for learning the multiscale information, and then the extracted multiscale features $X_{l+1}^L \in R^{3\tau_{l+1} \times w_{l+1} \times h_{l+1} \times c_{l+1}}$ are mapped into the category space by a GAP layer whose output is given by X_{GAP}^L . Particularly, l is set to 1 in our experiments.

Finally, X_{GAP}^L is fed to a softmax function to obtain the conditional probability distribution P_c^L , and the cross entropy is also set as the loss function $Loss_L$ of the LiDAR channel to yield the final classification results.

F. Three-Level Fusion Strategy

An effective fusion network (with a three-level fusion strategy) is designed for making full use of the complementarity of the HSI and LiDAR data. This network is marked in Fig. 1 with red arrows.

In the first-level fusion, to fully exploit the (more complete) spatial information carried out by the LiDAR data to enhance the feature representation in the HSI branch, the spatial attention in SaAB is applied to the output of SeAB for composite attention learning. Furthermore, residual learning is also utilized to avoid the degradation problem. The forward propagation of this part is expressed as:

$$F_{\hat{X}_l^H} = \alpha_{Sa}^L \odot \hat{X}_l^H + \hat{X}_l^H, \tag{10}$$

where $F_{\hat{X}_{l}^{H}}$ denotes the spatial-spectral features of the HSI channel.

In the second-level fusion, the outputs of MSRAB in each branch are cascaded in the spectral dimension, and then, a 1×1 ConvLSTM3D layer and a GAP layer are utilized to fuse the cascaded features, as shown in Fig. 1. The forward propagation of this part is expressed as:

$$X_{GAP}^{F} = f_{GAP}(f_{CL3D1}([X_{l+1}^{H}, X_{l+1}^{L}])), \quad (11)$$

where X_{GAP}^F denotes the output of the GAP layer, and $f_{CL3D1}(\cdot)$ denotes the 1×1 ConvLSTM3D layer.

Finally, due to the fact that the order of magnitude of the HSI features is much larger than that of the LiDAR features, the impact of the LiDAR channel on the classification performance may need to be upscaled. Hence, at the top of the designed fusion network, the LiDAR features X_{GAP}^L are reused by cascading them with the features in (11), i.e., the third-level fusion. The outputs of this part are written as $X^F = [X_{GAP}^F, X_{GAP}^L]$, and then fed into a softmax function to yield the probability distribution P_c^F . Similar to the HSI and LiDAR branches, the cross entropy is still used as the loss function $Loss_F$ to optimize the fusion network.

G. Loss Function and Network Training Strategy

Given the loss functions in subsections III-D, III-E and III-F, a multi-task loss function for optimizing the proposed dual-channel A^{3} CLNN model is designed as:

$$Loss = \alpha Loss_H + \beta Loss_L + \gamma Loss_F, \qquad (12)$$

where α , β , and γ are the scalar weights. For convenience, they are fixed to 1 in our experiments.

All the weights and biases in the proposed dualchannel A^3 CLNN model need to be learned. In order to train the whole model adequately –different from the training strategies in [34]– a novel and effective stepwise training approach is proposed. In the first stage, the LiDAR and HSI branches are trained using N_{step1} and N_{step2} epochs, respectively, which can provide the primary spatial features and spatial-spectral features to the designed fusion network and be regarded as pretrained channels to replace the ones obtained by traditional random initialization. Then, inspired by transfer learning [53], the proposed fusion network is initialized by these two pre-trained branches, and the multi-task loss function in (12) is further optimized in N_{steps} epochs to yield the final classification results of the proposed model. This also accelerates the convergence speed of the whole model. The detailed training strategy is summarized in Algorithm 1.

Algorithm 1 '	Fraining Dual-Channe	$1 A^{3}$ CLNN for 1	Mul-
tisource Remo	te Sensing Data Class	sification	

- **Input:** HSI data X^H ; The LiDAR data X^L ; Ground truth Y **Output:** Classification map Ω
- 1: Parameter setting and weights initialization
- 2: while step $\leq N_{step1}$ do
- 3: Train the LiDAR branch by optimizing the loss function $Loss_L$
- 4: Save model as the pre-trained LiDAR model
- 5: end while
- 6: while step $\leq N_{step2}$ do Train the HSI branch by optimizing the loss function $Loss_H$
- 7: Save model as the pre-trained HSI model
- 8: end while
- 9: while step $\leq N_{steps}$ do Restore these two pre-trained models to initialize the fusion network
- 10: Train the whole classification model by optimizing the multi-task loss function Loss
- 11: end while
- 12: return Classification map Ω

It should be pointed out that the adaptive momentum (ADAM) algorithm is adopted to optimize the three loss functions $Loss_H$, $Loss_L$, and Loss with different learning rates, which are represented by lr_H , lr_L , and lr, respectively. Additional explanations on parameter settings will be given in Section IV.

IV. EXPERIMENTAL RESULTS

In order to quantitatively and qualitatively evaluate the performance of the proposed dual-channel A^3 CLNN model, some state-of-the-art methods are selected for comparison, such as ELM [11], SVM [34], SSCL3DNN and SaCL2DNN [30], two-branch CNN [34], and dualchannel CapsNet [36]. The overall (OA), average accuracy (AA), and Kappa coefficient (κ) are utilized as the quantitative metrics to measure the classification performance of all algorithms. For the sake of eliminating the bias caused by random initialization of parameters in deep learning-based models, all experiments are repeated 10 times, and the average value is given for each quantitative metric. All our experiments have been conducted on a desktop PC with an Intel Core i7-8700 processor and an Nvidia GeForce GTX 1080ti GPU.



Fig. 5. (First row) false-color map (using bands 57, 27 and 17) of the Houston HSI data. (Second row) grayscale representation of the Houston LiDAR data. (Third row) ground-truth map of the Houston data set.

A. Experimental Data

In our experiments, two HSI + LiDAR data sets, i.e., Houston data set and Trento data set, are considered to evaluate the performance of our dual-channel A^{3} CLNN. According to [9] and [32], the false-color maps, grayscale representations, ground-truth maps, and the training samples for each data set are respectively presented in Figs. 5-6 and Tables I-II. In the following, we describe these data sets in more detail:



Fig. 6. (First row) false-color map (using bands 25, 15 and 2) of the Trento HSI data. (Second row) grayscale representation of the Trento LiDAR data. (Third row) ground-truth map of the Trento data set.

1) Houston Data: These data were captured in 2012 by the Compact Airborne Spectrographic Imager (CASI) sensor over the University of Houston campus and the surrounding area. The data was introduced in the 2013 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion contest. Its size is 349×1905 pixels, with spatial resolution of 2.5 m. There are 144 bands in the wavelength range from 0.38 to 1.05 μ m and 15 distinguishable class labels. These data, including the reference classes, are available online from the IEEE GRSS Data and Algorithm Standard Evaluation Website: http://dase.grss-ieee.org/.

 TABLE I

 Number of Training Samples for The Houston Data

NO.	Color	Class	Training	Total
1		Health grass	198	1251
2		Stressed grass	190	1254
3		Synthetic grass	192	697
4		Tress	188	1244
5		Soil	186	1242
6		Water	182	325
7		Residential	196	1268
8		Commercial	191	1244
9		Road	193	1252
10		Highway	191	1227
11		Railway	181	1235
12		Parking lot 1	192	1233
13		Parking lot 2	184	469
14		Tennis court	181	428
15		Running track	187	660
		U U		
Total			2832	15029

 TABLE II

 Number of Training Samples for The Trento Data



2) *Trento Data:* These data were collected by the AISA Eagle sensor over a rural area in Trento, Italy. The

data comprises 600×166 pixels with spatial resolution of 1 m, 6 ground-truth classes, and 63 bands in the spectral range from 420.89 to 989.09 nm.



Fig. 7. Overall accuracy (%) achieved by the proposed dual-channel A^3 CLNN model with different parameters for the Houston and Trento data sets. (a) Size $s \times s$ of the local spatial window. (b) Number K of principal components. (c)-(d) Learning rates lr_H and lr_L . (e)-(f) Learning rate lr. (g)-(h) Training epochs N_{step1} and N_{step2} .

B. Parameter Settings

As in [15] and [30], PCA is also utilized as the dimension reduction approach. For the HSI branch in the dual-channel A^3 CLNN model, the first K components after PCA are retained as the spectral information.

For the compared algorithms, the parameter settings of SVM, ELM, SSCL3DNN, SaCL2DNN, two-branch CNN, and dual-channel CapsNet are obtained according to [34], [11], [30], [34], [36] to achieve quasi-optimal performance. For the proposed dual-channel A^3 CLNN model in Fig. 1, there are some parameters that need to be tuned, i.e., the size ($s \times s$) of the local spatial window, the number (K) of the principal components, the size ($k \times k$) of the convolution kernels, the number (M) of feature maps, the value of the dropout operation, the

learning rates $(lr_H, lr_L, and lr)$, and the training epochs $(N_{step1} \text{ and } N_{step2})$. At first, K is fixed to 10, and the value of the dropout operation is 0.5. The learning rates $[lr_H, lr_L, lr]$ for the Houston data set are set to [0.0001, 0.001, 0.0001], respectively, and to [0.001, 0.0005, 0.0001] for the Trento data set, respectively. The training epochs N_{step1} and N_{step2} for the two data sets are fixed to 500. Parameter M in the first layer of each branch, the $a \times a$ ConvLSTM layer of MSRAB, and the ConvLSTM3D layer of the fusion network is respectively fixed to $\{32, 64, 128\}$. The number N_{steps} of training epochs is fixed to 1200 for the two data sets. After that, s is searched from $\{9, 11, 13, 15\}$ for the two data sets, and k is set to a value in the range $\{3, 4, 5\}$. Based on the above parameter settings, the experimental results for analyzing the effect of using different spatial window sizes on the classification performance are reported in Fig. 7(a), from which it is obvious that the optimal size of the local window is set to 13×13 (Houston data) and 11×11 (Trento data), respectively.

Furthermore, an optimal number K is generated from a given set $\{5, 10, 15, 20\}$, and the OA achieved by the proposed method (for different values of parameter K) is shown in Fig. 7(b), from which it can be seen that the optimal value of K is 10 for the two data sets.

Then, the performance on different values of M for four different combinations: $\{8, 16, 32\}$, $\{16, 32, 64\}$, $\{32, 64, 128\}$, and $\{64, 128, 256\}$ is analyzed in Table III. From Table III, we can infer that the optimal number of feature maps is $\{32, 64, 128\}$ (Houston data) and $\{16, 32, 64\}$ (Trento data), respectively.

To reduce the occurrence of overfitting problem, the dropout operation is used as a training trick, and the experiments for analyzing the influence of its different values on the classification performance are conducted. From Table IV, the optimal value of the dropout operation is set to 0.5 for the two data sets.

The learning rate is one of the parameters that controls the convergence rate of the proposed dual-channel A^3 CLNN model in the training process, and the optimal values of lr_H , lr_L and lr are searched in the range {0.0001, 0.0005, 0.001, 0.005, 0.01}. Concretely, Figs. 7(c)-(d) report the experimental results obtained after tuning the learning rates when lr was fixed to 0.0001. From Figs. 7(c)-(d), it is evident that the optimal value of $[lr_H, lr_L]$ for the two data sets is [0.0001, 0.001] (Houston data) and [0.001, 0.0005] (Trento data), respectively. Furthermore, the experimental results obtained after tuning lr are shown in Figs. 7(e)-(f), from which it can be seen that the learning rate lr for the two data sets can be fixed to 0.0001.

TABLE III SENSITIVITY COMPARISON AND ANALYSIS OF THE FEATURE MAPS OBTAINED FOR DIFFERENT VALUES OF M

M	Houston Data Set	Trento Data Set
$\{8, 16, 32\}$	85.96	97.56
$\{16, 32, 64\}$	87.87	98.54
$\{32, 64, 128\}$	90.82	98.27
$\{64, 128, 256\}$	88.64	96.35

TABLE IV Sensitivity Comparison and Analysis of Different Values of The Dropout Operation

Dropout	Houston Data Set	Trento Data Set
0.4	89.42	98.22
0.5	90.82	98.54
0.6	88.69	98.14

 TABLE V

 Parameter Settings for The Houston Data Set

Layer Name	Kernel Size	Output Size for HSI Channel	Output Size for LiDAR Channel
Input		$13 \times 13 \times 10 \times 1$	$13 \times 13 \times 1$
ConvLSTM3D	$3 \times 3 \times 3$	$13 \times 13 \times 10 \times 32$	
ConvLSTM2D	3×3		$13 \times 13 \times 32$
SeAB		$13 \times 13 \times 10 \times 32$	
SaAB			$13\times13\times32$
First-Level Fusion		$13 \times 13 \times 13$	10×32
MaxPooling3D	$2 \times 2 \times 2$	$7 \times 7 \times 5 \times 32$	
MaxPooling2D	2×2		$7 \times 7 \times 32$
MSRAB(3D)	$4 \times 4 \times 4$	$3 \times 7 \times 7 \times 5 \times 64$	
MSRAB(2D)	3×3		$3 \times 7 \times 7 \times 64$
Second-Level Fusion	$1 \times 1 \times 1$	$3 \times 7 \times 7 \times$	6×128
Dropout	0.5		
GAP3D	$3 \times 7 \times 7 \times 6$	$1 \times 1 \times 1 \times$	1×128
Dropout	0.5		
GAP3D	$3 \times 7 \times 7 \times 5$	$1 \times 1 \times 1 \times 1 \times 64$	
GAP2D	$3 \times 7 \times 7$		$1 \times 1 \times 1 \times 64$
Third-Level Fusion		$1 \times 1 \times 1 \times$	1×192
Dropout	0.5		
Softmax		13	13

 TABLE VI

 Parameter Settings for The Trento Data Set

Layer Name	Kernel Size	Output Size for HSI Channel	Output Size for LiDAR Channel
Input		$11 \times 11 \times 10 \times 1$	$11 \times 11 \times 2$
ConvLSTM3D	$3 \times 3 \times 3$	$11 \times 11 \times 10 \times 16$	
ConvLSTM2D	3×3		$11 \times 11 \times 16$
SeAB		$11 \times 11 \times 10 \times 16$	
SaAB			$11\times11\times16$
First-Level Fusion		$11 \times 11 \times 11$	10×16
MaxPooling3D	$2 \times 2 \times 2$	$6 \times 6 \times 5 \times 16$	
MaxPooling2D	2×2		$6 \times 6 \times 16$
MSRAB(3D)	$4 \times 4 \times 4$	$3 \times 6 \times 6 \times 5 \times 32$	
MSRAB(2D)	3×3		$3 \times 6 \times 6 \times 32$
Second-Level Fusion	$1 \times 1 \times 1$	$3 \times 6 \times 6 \times$	6×64
Dropout	0.5		
GAP3D	$3 \times 6 \times 6 \times 6$	$1 \times 1 \times 1 \times$	1×64
Dropout	0.5		
GAP3D	$3 \times 6 \times 6 \times 5$	$1\times1\times1\times1\times32$	
GAP2D	$3 \times 6 \times 6$		$1 \times 1 \times 1 \times 32$
Third-Level Fusion		$1 \times 1 \times 1 \times$	1×96
Dropout	0.5		
Softmax		13	13

Finally, for the training epochs N_{step1} and N_{step2} , we carry out the experiments to study the effect of

the training epochs of these two pre-training channels, and the optimal N_{step1} and N_{step2} are selected from $\{400, 500, 600\}$. As shown in Figs. 7(g)-(h), the optimal $[N_{step1}, N_{step2}]$ for the two data sets is [500, 500]. The detailed parameter settings for the proposed model for the two data sets are reported in Tables V-VI.

C. Classification Performance

According to [9] and [32], we evaluate the performance of the considered classification algorithms on the Houston and Trento data sets, using the available training samples. The obtained results are reported in Tables I-II. Note that the data enhancement technology in [34] is utilized to extend training sets for two-branch CNN and dual-channel CapsNet.

TABLE VII Classification Performance of Each Branch: HSI (H) AND LIDAR (L)

Data Sat	Propo	sed (L)	Propose	ed (H)	Proposed	1 (H+L)
Data Set	OA	Kappa	OA Kappa		OA Kappa	
Houston Data	59.83	56.48	87.00	85.90	90.55	89.75
Trento Data	89.31	85.98	97.65	96.86	98.73	98.31

On the basis of the parameter settings reported in subsection IV-B, the experimental results obtained by all the considered methods on the HSI data alone, the LiDAR data alone, and HSI + LiDAR data are reported in Tables VII-X. For convenience, HSI data, LiDAR data, and HSI + LiDAR data in Tables VII-X are abbreviated as H, L, and H+L, respectively. Particularly, for SSCL3DNN (which merges H+L) in Tables VIII-IX, the input of SSCL3DNN [30] is changed to the fusion of the HSI and LiDAR data, and the SaCL2DNN model [30] is selected for extracting spatial features from the LiDAR data, since SSCL3DNN is not suitable for dealing with 2-D data (and also because SaCL2DNN has a similar structure with regards to SSCL3DNN). From Tables VII-X, it can be seen that the proposed dual-channel A^{3} CLNN can obtain better classification performance than the other tested methods. On the one hand, the gate mechanisms realized by the convolution operation make it possible for the ConvLSTM2D and ConvLSTM3D layers to fully exploit both spatial and spectral information from multisource remote sensing data. On the other hand, with the help of the spectral, spatial and multiscale attention mechanisms, the proposed model can extract highly effective spectral- and spatial-enhanced features, and fully exploit multiscale information coming from multisource remote sensing data. Furthermore, the threelevel fusion and stepwise training strategies not only can fully integrate the spectral and spatial information by exploiting the complementary information of HSI and LiDAR data, but also accelerate the convergence speed of the whole model. Concretely, the experimental

 TABLE VIII

 Classification Results Achieved by Different Approaches for The Houston Data Set

	SVM	SVM	FIM	FIM	Two Branch	Two Branch	Dual Channel	Dual Channel	SSCI 3DNN	SSCI 3DNN	Proposed	Proposed
Class			LLM		Two-Branch	CNDV(LLL)			SSCLEDININ	SSCL5DINN	rioposed	rioposed
	(H)	(H+L)	(H)	(H+L)	CNN(H)	CNN(H+L)	CapsNet(H)	CapsNet(H+L)	(H)	(Merge, H+L)	(H)	(H+L)
1	81.01	82.53	82.15	82.24	94.85	97.98	80.63	81.39	81.04	82.05	79.84	81.73
2	82.24	84.77	82.93	82.99	82.59	89.44	81.95	83.08	84.21	80.98	85.15	84.43
3	82.97	86.93	95.45	96.96	30.32	53.20	94.46	97.43	72.21	89.44	93.20	91.49
4	90.81	95.83	90.44	91.41	98.30	96.73	90.06	88.64	90.79	90.85	89.20	96.72
5	97.63	97.54	99.59	98.96	95.61	96.88	100.00	100.00	100.00	99.78	99.84	99.97
6	79.72	88.81	71.79	77.86	76.41	24.31	89.51	95.10	84.38	87.18	95.34	97.90
7	76.12	81.16	80.32	74.91	96.18	87.47	81.72	91.23	89.55	91.51	84.58	87.06
8	43.40	44.92	64.10	63.60	70.54	89.02	71.51	92.40	68.66	93.32	81.83	96.93
9	79.41	86.40	72.62	77.34	82.31	86.85	72.24	80.64	87.91	78.88	86.02	87.88
10	90.15	59.75	80.79	58.91	65.41	78.20	62.26	65.54	52.38	55.60	60.42	70.82
11	63.00	71.82	68.22	88.58	78.69	90.88	72.87	88.99	73.62	90.83	95.70	98.13
12	84.15	92.41	72.81	78.87	83.75	65.99	86.55	87.42	93.05	91.80	93.05	94.65
13	89.82	85.96	42.57	54.97	94.25	100.00	77.89	62.46	92.05	85.96	91.46	96.02
14	80.97	83.00	90.01	92.44	95.02	98.28	93.93	95.95	92.31	78.41	99.60	97.30
15	66.60	74.21	84.00	93.94	91.10	96.37	94.93	96.41	94.43	94.86	99.86	96.05
OA	78.79	80.15	79.52	80.76	77.79	83.15	81.53	86.61	82.72	86.01	87.00	90.55
AA	79.20	81.07	78.52	80.96	82.35	83.44	83.37	87.11	83.79	86.10	89.01	91.81
κ	77.15	78.58	77.74	79.10	75.95	81.73	80.01	85.50	81.33	84.84	85.90	89.75

TABLE IX Classification Results Achieved by Different Approaches for The Trento Data Set

Class	SVM	SVM	ELM	ELM	Two-Branch	Two-Branch	Dual-Channel	Dual-Channel	SSCL3DNN	SSCL3DNN	Proposed	Proposed
Class	(H)	(H+L)	(H)	(H+L)	CNN(H)	CNN(H+L)	CapsNet(H)	CapsNet(H+L)	(H)	(Merge, H+L)	(H)	(H+L)
1	59.59	97.69	89.31	93.17	90.38	97.44	98.46	97.15	96.48	98.32	98.17	98.92
2	34.55	87.36	71.55	87.95	96.66	93.29	94.14	99.07	91.62	96.88	95.97	99.14
3	92.69	87.06	92.21	73.56	88.24	72.86	91.44	97.29	90.81	82.19	91.79	98.12
4	98.61	99.80	97.58	97.30	99.32	98.01	96.17	100.00	97.15	99.81	99.59	100.00
5	97.93	93.36	87.57	93.17	98.24	98.44	98.93	94.62	99.87	96.74	99.89	99.95
6	79.30	69.34	59.84	66.95	60.09	80.14	72.21	91.71	79.31	85.34	86.40	90.57
OA	84.89	92.69	86.45	90.85	93.20	95.36	94.65	96.75	95.50	96.46	97.65	98.73
AA	77.11	89.10	83.01	85.35	88.82	90.03	91.89	96.64	92.54	93.21	95.30	97.78
κ	79.45	90.22	82.01	87.81	90.92	93.81	92.87	95.69	93.99	95.30	96.86	98.31



Fig. 8. Classification maps obtained by different approaches for the Houston data set. (a) SVM (80.15%). (b) ELM (80.76%). (c) Two-branch CNN (83.15%). (d) dual-channel CapsNet (86.61%). (e) SSCL3DNN (86.01%). (f) The proposed dual-channel A³CLNN (90.55%).



Fig. 9. Classification maps obtained by different approaches for the Trento data set. (a) SVM (92.69%). (b) ELM (90.85%). (c) Two-branch CNN (95.36%). (d) dual-channel CapsNet (96.75%). (e) SSCL3DNN (96.46%). (f) The proposed dual-channel A³CLNN (98.73%).

TABLE X
CLASSIFICATION RESULTS OBTAINED BY DIFFERENT APPROACHES FOR THE LIDAR DATA OF THE TWO CONSIDERED DATA SETS

			He	ouston Data Set					Ti	rento Data Set		
Class	SVM	ELM	Two-Branch	Dual-Channel	SaCL2DNN	Proposed	SVM	ELM	Two-Branch	Dual-Channel	SaCL2DNN	Proposed
	(L)	(L)	CNN(L)	CapsNet(L)	(L)	(L)	(L)	(L)	CNN(L)	CapsNet(L)	(L)	(L)
1	20.23	7.41	39.61	43.87	58.97	52.30	37.23	11.38	87.30	90.31	91.41	88.84
2	15.51	2.88	22.03	26.69	26.00	35.40	61.73	68.27	88.93	91.35	95.64	89.34
3	40.20	26.40	68.37	82.57	28.98	46.60	50.73	20.04	62.00	72.65	57.41	73.76
4	94.98	37.69	81.81	66.48	74.59	79.83	69.18	66.58	99.28	93.23	97.66	94.25
5	26.42	13.79	60.74	38.83	24.87	44.92	28.80	47.91	65.32	83.07	68.61	86.93
6	60.14	55.94	21.40	31.47	37.53	52.91	71.80	24.20	81.72	61.85	84.95	85.90
7	41.60	38.62	80.15	55.50	88.22	79.20						
8	65.81	75.21	65.96	86.80	75.34	92.47						
9	11.80	11.49	60.90	40.42	54.04	51.15						
10	9.74	7.01	41.54	50.29	58.72	46.07						
11	35.58	35.96	74.20	66.70	82.64	89.25						
12	12.10	3.33	41.62	42.75	13.74	38.52						
13	31.23	47.49	49.70	51.23	74.50	61.29						
14	76.92	27.6	39.24	85.83	53.85	54.12						
15	3.81	14.59	39.81	67.86	20.79	53.63						
OA	33.71	24.21	53.61	54.15	53.50	59.83	50.15	47.69	82.45	85.50	84.56	89.31
AA	36.41	27.03	52.47	55.82	51.52	58.49	53.24	39.73	80.76	82.08	82.62	86.51
κ	28.63	19.26	49.80	50.55	49.68	56.48	38.48	32.28	77.24	80.99	79.98	85.98

results obtained when analyzing the effectiveness of the proposed three-level fusion strategy and stepwise training strategy are given in Table VII, from which it is apparent that, for the LiDAR data in Houston and Trento data sets, the values of the OA metric in the LiDAR branch are 59.83% and 89.31%, respectively, while the values of the OA metric in the HSI branch are 87.00% and 97.65%, respectively. Moreover, when utilizing HSI and LiDAR data, the gains in OA obtained by using our newly developed model are respectively 3.55% and 1.08% (compared to the HSI channel) for the two considered data sets. This demonstrates the effectiveness of the proposed dual-channel A^3 CLNN.

Moveover, the experimental results aimed at analyzing the accuracy obtained for each class and quantitative metrics obtained after fusing HSI and LiDAR data are reported in Tables VIII-X. From these tables it can be seen that, without data augmentation, the proposed dual-channel A^{3} CLNN model performs better than other baseline methods. Concretely, for the Houston data set, the proposed model yields highly competitive classification accuracy of 90.55%, with a gain over 7.40% with respect to that achieved by the two-branch CNN. Our model also yields 10.40% and 9.79% improvements with regards to the standard SVM and ELM, respectively. As for the Trento data set, the improvements in OA achieved by the proposed dual-channel A^{3} CLNN model are 6.04%, 7.88%, and 3.37%, respectively, when compared with SVM, ELM, and two-branch CNN. For dual-channel CapsNet [36], although many works have verified that CapsNet can better learn the information of position, orientation, deformation, and texture than CNN, the spectral and scale information contained in different classes may not be fully learned. The proposed model outperforms dual-channel CapsNet, having improvements of 3.94% and 1.98% for the two data sets, respectively. The above experimental results show that, for traditional models such as SVM and ELM, the way of converting the HSI and LiDAR data into vectors leads to the loss of the spatial and geometric structure information, while simple feature cascading in two-branch CNN [34] and dual-channel CapsNet [36] fails to effectively learn complementary information by fusing the HSI and LiDAR data. In contrast to these comparison algorithms, the design of gate mechanisms makes the ConvLSTM-based models to fully leverage spatial information and better preserve the intrinsic structure information of the original data, which is in line with the findings in [30]. In addition, compared with SSCL3DNN, our dual-channel A³CLNN model can improve the classification accuracy in almost all cases, and obtain 4.54% and 2.27% gains in OA for the two data sets, respectively, benefiting from the developed threelevel fusion and stepwise training strategies. In addition, when only HSI or LiDAR data are used, the performance of each branch in the proposed model is also superior to that achieved by other comparison methods. Specifically, the dual-channel A^{3} CLNN(H) model can even achieve higher accuracy than other models that use both HSI and LiDAR data, which may be the contributions of the spectral and multiscale attention structures. More detailed experimental results can be found in Tables VIII-X.

Regarding Tables VIII-IX, similar conclusions can be drawn from the classification maps presented in Figs. 8-9, from which it is obvious that the classification maps yielded by the proposed dual-channel A^{3} CLNN model are the closest to the ground-truth maps for the two considered data sets. Particularly, the obtained classification maps exhibit less mislabeled areas, and the boundaries between different classes are better delineated and identified, especially for classes 6, 8 and 11 in Fig. 8, and classes 2, 3 and 5 in Fig. 9. This further verifies the advantages and effectiveness of our dualchannel A^3 CLNN model.

D. Ablation Study

To highlight the effectiveness of SeAB, SaAB, MSRAB, two-level attention strategy, and stepwise training strategy in dual-channel A^3 CLNN, detailed ablation studies are conducted to see how they contribute to the classification performance. SSCL3DNN [30] is used as a baseline. For convenience, the dual-channel A^3 CLNN model with and without each component are respectively abbreviated as proposed(\cdot , with) and proposed(\cdot , without), in which \cdot denotes H, L, and H+L.

(1) The effectiveness of SeAB: To effectively learn the spectral-enhanced feature representation from the HSI data, a SeAB module is built for the classification of the HSI branch. To demonstrate its effectiveness, we conduct experiments to analyze the influence of SeAB by adding and removing it from our dual-channel A^{3} CLNN, whose results are reported in Table XI. Compared with SSCL3DNN(H), the proposed model(H, without SeAB) can obtain 3.13% and 1.57% gains for the two data sets, respectively, while the proposed model(H+L, without SeAB) generates 4.04% and 1.65% improvements against SSCL3DNN(H+L). With the help of SeAB, the proposed model(H+L) further improves the classification accuracy of the proposed model(H+L, without SeAB) by 0.50% and 0.62% for the two data sets, indicating that the SeAB module can improve the classification performance by enhancing the ability of spectral feature representation of the whole model.

(2) The structure analysis of SaAB: LiDAR data can provide rich elevation information in the spatial domain, thus having the potential of improving the characterization of HSI scenes. The experimental results for analyzing the performance of our dual-channel A^3 CLNN model with and without SaAB are reported in Table XII. Compared with the proposed model(H+L, without SaAB), 1.22% and 0.56% gains are yielded by the proposed model(H+L, with SaAB) for the two data sets, respectively, which shows the advantages of SaAB.

(3) The analysis of two-level attention strategy: To fully utilize and fuse the spectral and spatial information, a composite attention learning module is designed as a two-level attention strategy to jointly learn spectral- and spatial-enhanced features. The experiments for analyzing its influence on the classification performance are carried out, whose results are presented in Table XIII. Compared with SSCL3DNN(H+L), the gains in OA yielded by the proposed model(H+L, without composite attention learning) are respectively 2.99% and 1.17% for the two

data sets, showing the superiority of the MSRAB model to improve the classification performance to some extent. Furthermore, after embedding the composite attention learning module, the proposed model(H+L) achieves 1.55% and 1.10% gains for the two data sets, respectively. Experimental results in Table XIII show that the composite attention learning and MSRAB modules can effectively learn the spatial-spectral and multiscale information, resulting in better classification results.

TABLE XI THE INFLUENCE OF THE PROPOSED SEAB MODULE

M- 1-1-	Houst	on Data	Trento	Data
Models	OA	Kappa	OA	Kappa
SSCL3DNN(H)	82.72	81.33	95.50	93.99
Proposed(H, without)	85.85	84.64	97.07	96.10
Proposed(H, with)	87.00	85.90	97.65	96.86
SSCL3DNN(H+L)	86.01	84.84	96.46	95.30
Proposed(H+L, without)	90.05	89.22	98.11	97.47
Proposed(H+L, with)	90.55	89.75	98.73	98.31

 TABLE XII

 The Structure Analysis of The Developed SAAB Module

Models	Houston Data		Trento Data	
	OA	Kappa	OA	Kappa
SaCL2DNN(L)	53.50	49.68	84.56	79.98
Proposed(L, without)	58.48	55.05	87.59	83.60
Proposed(L, with)	59.83	56.48	89.31	85.98
SSCL3DNN(H+L)	86.01	84.84	96.46	95.30
Proposed(H+L, without)	89.33	88.45	98.17	97.56
Proposed(H+L, with)	90.55	89.75	98.73	98.31

 TABLE XIII

 The Analysis of Two-Level Attention Strategy

Models	Houston Data		Trento Data	
	OA	Kappa	OA	Kappa
SSCL3DNN(H+L)	86.01	84.84	96.46	95.30
Proposed(H+L, without)	89.00	88.11	97.63	96.83
Proposed(H+L, with)	90.55	89.75	98.73	98.31

TABLE XIV THE ARCHITECTURE STUDY OF THE DESIGNED MSRAB MODULE

Models	Houston Data		Trento Data	
	OA	Kappa	OA	Kappa
SSCL3DNN(H)	82.72	81.33	95.50	93.99
Proposed(H, without)	85.29	84.08	96.33	95.11
Proposed(H, with)	87.00	85.90	97.65	96.86
SaCL2DNN(L)	53.50	49.68	84.56	79.98
Proposed(L, without)	56.34	52.73	87.81	84.16
Proposed(L, with)	59.83	56.48	89.31	85.98
SSCL3DNN(H+L)	86.01	84.84	96.46	95.30
Proposed(H+L, without)	87.23	86.23	97.58	96.77
Proposed(H+L, with)	90.55	89.75	98.73	98.31

(4) The architecture study of MSRAB: Different classes in hyperspectral data should contain different scale information, so that using the fixed-scale convolution kernel limits the ability of CNN-based models to learn scale information. To analyze the contributions of MSRAB, we compare the proposed model with a variant of our model in which the MSRAB module is replaced

with a ConvLSTM2D or ConvLSTM3D layer in Table XIV. We can observe that, compared with this variant, MSRAB can bring 3.32% and 1.15% improvements to the proposed model(H+L) for the two data sets, respectively. Moreover, the experimental results shown in the last three lines of Table XIV also validate that the composite attention learning and MSRAB modules are one of the main reasons for the performance improvements of dual-channel A^3 CLNN, which is consistent with the conclusions in Table XIII. More detailed experimental results are reported in Table XIV.



Fig. 10. The learning curves of the proposed dual-channel A^3 CLNN for the Houston and Trento data sets. (a) The learning curves for the Houston data set. (b) The learning curves for the Trento data set.

(5) The analysis of stepwise training strategy: To obtain better optimization of our dual-channel A³CLNN, a stepwise training strategy is designed for the full training of the whole model. The learning curves of the proposed method on the two data sets are visualized in Fig. 10, where both curves for each data set tend to converge stably with the increase of the number of iterations. By analyzing the data distribution of the two data sets in Figs. 5-6, we can find that the data distribution of the Houston data set is more scattered than that of the Trento data set. From Fig. 10(a), although these two curves fluctuate slightly, the loss curve of the proposed model with stepwise training strategy converges faster and is more stable than the one without it. The learning curves shown in Fig. 10 illustrate that the proposed stepwise training strategy is advantageous for accelerating the convergence speed of the proposed dual-channel A^{3} CLNN model to some extent.

V. CONCLUSION

In this paper, a new dual-channel A^3 CLNN model has been proposed for the classification of multisource remote sensing data. Specifically, our model comprises two different pipelines for LiDAR data and HSI data, in which spatial, spectral, and multiscale residual attention structures have been implemented to fully exploit spatial, spectral, and multiscale information and obtain more comprehensive and discriminative feature representation. Moreover, an effective three-level fusion strategy and a novel stepwise training strategy are also developed to fully integrate the spatial and spectral information contained in the LiDAR and HSI data, exploiting their complementarity. Our experimental results demonstrate that the proposed dual-channel A^{3} CLNN provides better performance than state-of-the-art CNN-based approaches (e.g., the two-branch CNN), the capsule network-based models (e.g., the dual-channel CapsNet) and baseline ConvLSTM-based methods (e.g., SSCL3DNN).

REFERENCES

- [1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6-36, Jun. 2013.
- [2] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45-54, Jan. 2014.
- [3] M. Belgiu, I. Tomljenovic, T. J. Lampoltshammer, T. Blaschke, and B. Höfle, "Ontology-based classification of building types detected from airborne laser scanning data," *Remote Sens.*, vol. 6, no. 2, pp. 1347-1366, Feb. 2014.
- [4] I. Tomljenovic, B. Höfle, D. Tiede, and T. Blaschke, "Building extraction from airborne laser scanning data: An analysis of the state of the art," *Remote Sens.*, vol. 7, no. 4, pp. 3826-3862, Apr. 2015.
- [5] J. Jung, E. Pasolli, S. Prasad, J. C. Tilton, and M. M. Crawford, "A framework for land cover classification using discrete return LiDAR data: Adopting pseudo-waveform and hierarchical segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 2, pp. 491-502, Feb. 2014.
- [6] W. Liao, R. Bellens, A. Pižurica, S. Gautama, and W. Philips, "Combining feature fusion and decision fusion for classification of hyperspectral and LiDAR data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Quebec City, QC, 2014, pp. 1241-1244.
- [7] M. Khodadadzadeh, A. Cuartero, J. Li, A. Felicsimo, and A. Plaza, "Fusion of hyperspectral and LiDAR data using generalized composite kernels: A case study in Extremadura, Spain," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Milan, Italy, 2015.
- [8] C. Zhao, X. Gao, Y. Wang, and J. Li, "Efficient multiple-feature learning-based hyperspectral image classification with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4052-4062, Jul. 2016.
- [9] B. Rasti, P. Ghamisi, and R. Gloaguen, "Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3997-4007, Jul. 2017.
- [10] L. Pan, H. Li, W. Li, X. Chen, G. Wu, and Q. Du, "Discriminant analysis of hyperspectral imagery using fast kernel sparse and lowrank graph," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6085-6098, Nov. 2017.
- [11] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681-3693, Jul. 2015.
- [12] Y. Deng, H. Li, L. Pan, L. Shao, Q. Du, and W. J. Emery, "Modified tensor locality preserving projection for dimensionality reduction of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 277-281, Feb. 2018.
- [13] M. Cui and S. Prasad, "Class-dependent sparse representation classifier for robust hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2683-2695, May 2015.
- [14] W. Hu, Y. Huang, W. Li, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, no. 2, pp. 1-12, Jul. 2015.
- [15] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232-6251, Oct. 2016.

- [16] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844-853, Feb. 2017.
 [17] W. Shao and S. Du, "Spectral-spatial feature extraction for
- [17] W. Shao and S. Du, "Spectral-spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544-4554, Aug. 2016.
- [18] B. Liu, X. Yu, P. Zhang, A. Yu, Q. Fu, and X. Wei, "Supervised deep feature extraction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 1909-1921, Apr. 2018.
- [19] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855-868, May 2009.
- [20] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639-3655, Jul. 2017.
- [21] X. Zhang, Y. Sun, K. Jiang, C. Li, L. Jiao, and H. Zhou, "Spatial sequential recurrent neural network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4141-4155, Nov. 2018.
- [22] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384-5394, Aug. 2019.
- [23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural comput.*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [24] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015.
- [25] H. Wu and S. Prasad, "Convolutional recurrent neural networks for hyperspectral data classification," *Remote Sens.*, vol. 9, no. 3, pp. 298, Mar. 2017.
- [26] A. Song, J. Choi, Y. Han, and Y. Kim, "Change detection in hyperspectral images using recurrent 3D fully convolutional networks," *Remote Sens.*, vol. 10, no. 11, pp. 1827, Nov. 2018.
- [27] M. Seydgar, A. A. Naeini, M. Zhang, W. Li, and M. Satari, "3-D convolution-recurrent networks for spectral-spatial classification of hyperspectral images," *Remote Sens.*, vol. 11, no. 7, pp. 883, Apr. 2019.
- [28] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Hyperspectral image classification using spectral-spatial LSTMs," *Neurocomputing*, vol. 328, no. 7, pp. 39-47, Feb. 2017.
- [29] Q. Liu, F. Zhou, R. Hang, and X. Yuan, "Bidirectionalconvolutional LSTM based spectral-spatial feature learning for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 12, pp. 1330, Dec. 2017.
- [30] W. Hu, H Li, L. Pan, W. Li, R. Tao, and Q. Du, "Spatial-spectral feature extraction via deep ConvLSTM neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4237-4250, Jun. 2020.
- [31] W. Liao, R. Bellens, A. Pizurica, S. Gautama, and W. Philips, "Graph-based feature fusion of hyperspectral and LiDAR remote sensing data using morphological features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Melbourne, Victoria, Australia, 2013.
- [32] M. Khodadadzadeh, J. Li, S. Prasad, and A. Plaza, "Fusion of hyperspectral and LiDAR remote sensing data using multiple feature learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2971-2983, Jun. 2015.
- [33] P. Ghamisi, R. Souza, J. A. Benediktsson, X. X. Zhu, L. Rittner, and R. A. Lotufo, "Extinction profiles for the classification of remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5631-5645, Oct. 2016.
- [34] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937-949, Feb. 2018.
- [35] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, "Feature extraction for classification of hyperspectral andLiDAR data using patch-to-patch cnn," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 100-111, Jan. 2020.

- [36] H. Li, W. Wang, L. Pan, W. Li, Q. Du and R. Tao, "Robust capsule network based on maximum correntropy criterion for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 738-751, 2020.
- [37] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996-1010, Apr. 2013.
- [38] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015.
- [39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000-6010.
- [40] L. Chen et al., "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, 2017, pp. 6298-6306.
- [41] H. Guo, K. Zhu, M. Tang, and J. Wang, "Two-level attention network with multi-grain ranking loss for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4328-4338, Sept. 2019.
- [42] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, 2017, pp. 6450-6458.
- [43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, 2018, pp. 7132-7141.
- [44] Q. Liu, X. Che, and M. Bie, "R-STAN: Residual spatial-temporal attention network for action recognition," *IEEE Access*, vol. 7, pp. 82246-82255, 2019.
- [45] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Trans. Med. Imaging*, vol. 38, no. 9, pp. 2092-2103, Sept. 2019.
- [46] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983-8997, Nov. 2019.
- [47] J. Chen, L. Wan, J. Zhu, G. Xu, and M. Deng, "Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 681-685, Apr. 2020.
- [48] C. Wang, X. Bai, S. Wang, J. Zhou, and P. Ren, "Multiscale visual attention networks for object detection in VHR remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 310-314, Feb. 2019.
- [49] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155-1167, Feb. 2019.
- [50] Z. Xiong, Y. Yuan, and Q. Wang, "AI-NET: Attention inception neural network for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Valencia, 2018.
- [51] L. Mou, Q. Liu, L. Bruzzone, and X. X. Zhu, "Learning spectralspatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924-935, Feb. 2019.
- [52] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065-8080, Oct. 2019.
- [53] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320-3328.
- [54] C. Szegedy et al, "Going deeper with convolutions," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, 2015, pp. 1-9.
- [55] Y. Tian, W. Hu, H. Jiang, and J. Wu, "Densely connected attentional pyramid residual network for human pose estimation," *Neurocomputing*, vol. 347, no. 28, pp. 13-23, Jun. 2019.
- [56] P. Ramachandran, B. Zoph, and Q. V. Le. (2018). "Searching for activation functions." [Online]. Available: https://arxiv.org/abs/1710.05941
- [57] M. Lin, Q. Chen, and S. Yan. (2014). "Network in network." [Online]. Available: https://arXiv.org/abs/1312.4400