Cross-Scene Building Identification Based on Dual-Stream Neural Network and Efficient Channel Attention Mechanism

Wenmei Li, *Member, IEEE*, Jiadong Zhang, Hao Xia, Qing Liu, Yu Wang, *Member, IEEE*, Yan Jia, and Yixiang Chen

***Abstract*—With the widespread popularity of deep learning, various neural network models are extensively employed in the recognition, classification, and segmentation of remote sensing images. Convolutional neural networks (CNNs), fully convolutional networks, including their variants like Unet, have demonstrated significant results within this particular domain. Nevertheless, CNNs exhibit limitations when it comes to grasping extended global dependencies. Conversely, transformers exhibit exceptional ability in effectively dealing with long-range dependencies. Considering this, we have introduced the efficient channel attention-enhanced dual-stream neural network (ECA-DSNN) to improve the identification of buildings across various scenes. Specifically, we developed a dual-stream network that incorporates the Unet and transformer framework in order to capture both the local and global context. In addition, we introduced an attention mechanism module to augment the model's generalization capability. With the advanced identification and generalization capability of ECA-DSNN, only fine-tuning and data augmentation are needed to achieve superior performance in cross-scene transfer, even with limited samples in the target domain. The outcomes indicated that the ECA-DSNN proposed achieved superior performance in comparison to the state-of-the-art methodologies, particularly in the experiment transferring from the source domain Beijing to the target domain Shanghai. In this scenario, the overall accuracy surpassed 96.3% and an F1 score exceeded 78.6%.**

***Index Terms*—Attention mechanism, building identification, cross-scene transfer, data enhancement transformer, limited samples, Unet.**

I. Introduction

THE interpretation of high-resolution remote sensing images is currently a focal point of research in the field of remote sensing processing. However, existing remote sensing information processing techniques lag behind the pace of remote sensing image acquisition, and there is a limited capacity to acquire and transfer knowledge between remote sensing images [1].

As spatial resolution increases, spectral heterogeneity also increases significantly. The phenomena of "same object presents a different spectrum" and "different object preserves the same spectrum" frequently occur. Moreover, the complexity of background in high-spatial resolution images leads to considerable interference in the extraction and identification of buildings. Given the characteristics of high-resolution images and buildings, the focus of research in remote sensing image interpretation lies in how to identify buildings from such images. Building identification (BI), a crucial component of high-definition remote sensing image interpretation [2], is widely employed in urban planning, population estimation, land utilization, and numerous other fields. Traditional BI methods rely on manually designed features (such as image element features, corner point features, spectral features, etc.) for extraction [3]. These methods are influenced by subjective factors, making it challenging to extract optimal features and imposing significant limitations. Before the widespread adoption of deep learning, approaches involving semiautomatic learning and active learning for extracting building footprint information from high-spatial resolution remote sensing images were proposed. Nevertheless, both active learning and semiautomatic learning excessively depend on expert knowledge, often overlook contextual information in images, and encounter challenges with limited annotated data that hinders generalization.

In recent years, deep learning has been widely applied in building recognition. By constructing numerous hidden layers and utilizing a substantial quantity of labeled data, deep learning

Manuscript received 13 October 2023; revised 13 December 2023 and 29 January 2024; accepted 4 March 2024. Date of publication 14 March 2024; date of current version 22 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42071414, in part by the Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of the People's Republic of China under Grant KLSMNR-K202201, in part by the Open Fund of State Key Laboratory of Remote Sensing Science under Grant OFSLRSS202202, in part by the Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of the People's Republic of China under Grant 202305, in part by the National Natural Science Foundation of China under Grant 41501378, in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant KYCX22_0977, and in part by the China Postdoctoral Science Foundation under Grant 2019M661896. *(Corresponding author: Yixiang Chen.)*

Wenmei Li, Yan Jia, and Yixiang Chen are with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210023, China, and also with the Health Big Data Analysis and Location Services Engineering Laboratory of Jiangsu Province, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: liwm@njupt.edu.cn; jiayan@njupt.edu.cn; chenyixiang@njupt.edu.cn).

Jiadong Zhang, Hao Xia, and Qing Liu are with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: 1022173208@njupt.edu.cn; 1023172909@njupt.edu.cn; 1021173512@njupt.edu.cn).

Yu Wang is with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: 1018010407@njupt.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3375321

efficiently achieves a greater understanding of building features for the purpose of building zone recognition. When compared to traditional manual feature extraction methods, deep learning excels at extracting profound building features, showcasing superior generalization ability in comparison to traditional building recognition. Convolutional neural networks (CNNs), a pivotal component of deep learning, are composed of multiple convolutional layers [4]. Unlike other deep learning architectures, CNNs are particularly well-suited for feature extraction and transfer, ensuring robust classification performance. The FCNs [5] along with its variations Unet and Resnet, which are developed based on CNNs, have made significant advancements in deep learning feature extraction. They are extensively employed in high-score image detection, scene classification, building footprint recognition, and other domains. The FCN utilizes an end-to-end network structure that harnesses image information comprehensively for feature extraction and hierarchical information modeling, leading to accurate predictions of buildings, in contrast to previous methods.

However, FCN and its variants are limited by the receptive field and demonstrate a lack of proficiency in handling spatial global context information. Many scholars have proposed diverse solutions to address this issue. These solutions primarily stem from two perspectives: One involves integrating the attention mechanism into the convolutional network to enhance global context understanding [6], while the other entails combining convolutional networks with transformers to convert 2-D data into a 1-D format [7], [8], facilitating robust sequence-to-sequence model predictions. For instance, the Swin-Transformer with sliding windows [9], A-GCRNN [10], and the TMUnet [11]. In BI tasks, a considerable quantity of labels from the source domain are adequate for classifying buildings within the same scene. Nonetheless, collecting a significant amount of labeled building data proves challenging, particularly when dealing with unlabeled data. With recently gathered data, achieving satisfactory accuracy remains difficult without a sufficiently large number of labels. Hence, cross-scene classification becomes imperative.

To address the issue of sparse annotation labels, cross-scene classification is proposed [12]. Transfer learning serves as a cross-scene classification method capable of addressing the challenge posed by insufficient labeled data [13]. In the scenario depicted in this article, both the source and target feature spaces align with each other, as well as the label space [14], giving rise to homogeneous migration learning. A relevant homogeneous transfer learning method was summarized by Wessi et al. in 2016 [15]. Transfer learning finds extensive application in object recognition, numerical classification, natural scene recognition, and similar domains [16]. Nonetheless, transfer learning encounters certain challenges due to the susceptibility of deep learning to noise. Ensuring the positivity of migrated features can be arduous, and there exists the possibility of negative feature migration. Embedding attentional mechanisms can help alleviate or even prevent the adverse effects of negative migration [17].

The attention mechanism has garnered significant popularity in deep learning due to its capacity to selectively focus on the most informative features while disregarding noise and irrelevant information [18]. In BI tasks, attention-based models have demonstrated promising outcomes by effectively capturing key building characteristics and enhancing recognition accuracy. By dynamically assigning distinct weights to input features, attention-based models have the capability to adaptively concentrate on the most pertinent image segments and incorporate highly informative details into the recognition process. Consequently, the attention mechanism has evolved into an indispensable tool for building recognition tasks, making significant contributions to the advancement of the field. As models grow in complexity, the volume of information stored within the model increases, potentially leading to information overload and feature selection bias.

To address this, we constructed a dual-stream feature extraction network based on the parallelism of Unet and Transformer. This approach compensates for the limitation of single-channel information extraction by employing dual information flow extraction. We improved the main feature information through feature fusion with multiple attention mechanisms to prevent noise and irrelevant information interference. Subsequently, we enhance building information in the target domain data using various data enhancement techniques to optimize data migration performance. Ultimately, we accomplish swift training and prediction in the target domain with minimal labeled data through migration fine-tuning. The primary contributions of this article are as follows.

1) We have designed a dual-stream neural network-supported framework for building identification, combining Transformer, and neural networks techniques for source domain feature learning.
2) We introduce an efficient channel attention (ECA)-based feature fusion module to achieve feature complementarity between different streams and improve classification performance.
3) We evaluate the ECA-DSNN method in cross-scene and verify its robustness and effectiveness. Through regularization, we enhance both shallow and primary features, thereby bolstering the model's feature learning capabilities. To address the issue of blurred outlines, we have modified common activation functions in CNNs to mitigate the appearance of jagged prediction outcomes. This modification partially rectifies the problem of blurred edges in building recognition.

II. Related Works

In recent years, three techniques have gained increasing popularity: transformer-based image recognition, attention mechanisms, and cross-scene object recognition migration. The subsequent three sections focus on the application of these techniques within the field of image processing.

A. Transformer-Based Image Recognition Methods

Traditional CNN model structures primarily emphasize local information aggregation, often struggling to effectively modeling long-range contextual information. With the emergence of transformers in the field of NLP, there is a growing interest

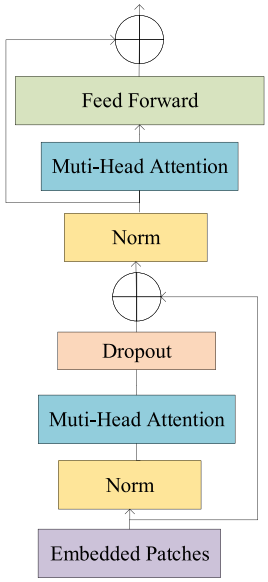

Fig. 1. Structure of ViT network encoder.

in its application to computer vision. The introduction of ViT marked a successful foray into applying transformer architecture from natural language to image-related challenges [7]. The transformer embodies a distinctive end-to-end network structure, encompassing two key components: an encoder and a decoder. The encoding facet incorporates multiple encoders, each comprising two sublayers: 1) an attention layer; and 2) a feed-forward network. Similarly, the decoding facet comprises several encoders, with each encoder layer composed of a self-attention layer and a feedforward network. Refer to Fig. 1 for an illustration depicting the structure, where each encoder contains a multiheaded attention mechanism and a feedforward network [19]. Numerous scholars have put forth a series of effective enhancement techniques to bolster the efficiency and effectiveness of ViT model training.

Han et al. [20] introduced the transformer in transformer model to improve model generalization and accuracy by refining each patch within the ViT framework. Nevertheless, while this model demonstrates improved performance on large-scale datasets, its results often remain unsatisfactory for small and medium-sized datasets. Yuan et al. [21] proposed the Tokens-to-Token VIT(T2T-ViT) model, which partially addresses this concern and achieves superior outcomes compared to the ViT model on medium-sized datasets. Liu et al. [9] pursued an approach of segmenting images into finer batches and progressively merging them at different layers, subsequently reducing resolution. This method enabled interaction among local information from distinct windows through the sliding window technique. However, the aforementioned methods only partially address the issue of the Transformer's overemphasis on global information, resulting in the neglect of local information. Enhancing the ViT model, as mentioned above, necessitates a substantial dataset; without it, achieving more efficient and accurate results compared to traditional building recognition methods becomes challenging.

B. Attention Mechanisms

Attentional mechanisms, fundamentally constituted by autonomously learned sets of weighting parameters within the network structure, employ "dynamic weighting" to highlight significant regions while reducing the influence of irrelevant background regions [22]. Recent strides in attentional mechanisms reveal substantial promise in the domains of computer vision and remote sensing. They offer avenues for accentuating meaningful data and diminishing the impact of distracting information, ultimately facilitating the extraction of representative features. In the realm of computer vision [23], [24], [25] and remote sensing image processing [26], [27], [28], [29], the attenuation of undesired information is pivotal for the extraction of representative features. The essence of these approaches can be summarized as follows.

1) Channel Attention Mechanism: It primarily focuses on interchannel relationships within the feature map, automatically determining the significance of each feature channel by learning channel-specific weights. These learned weights are subsequently utilized to amplify relevant features while suppressing less relevant ones for the current task. In contemporary times, the channel attention mechanism has found extensive application within the field of remote sensing images. For instance, Ge et al. [30] harnessed the channel attention mechanism to augment the retrieval capability of remote sensing images. In their article [31], SENet [23] was fused with the semantic pyramidal attention module to establish a global attention mechanism. This mechanism facilitates the extraction of high-level features and the enhancement of change-related information. Furthermore, variants like the ECANet attention mechanisms [32], enhanced by SENet attention mechanisms, have also emerged.

2) Spatial Attention Mechanism: The transformation of various spatial deformation data and the automatic capturing of crucial regional features ensure that an image can be cropped, panned, or rotated while retaining the original image's outcome after the operation [33]. Spatial attention mechanisms find widespread application in the realm of computer vision to recover more representative spatial objects. Currently, multiple approaches to spatial attention mechanisms exist, including spatial alteration neural networks [33], dynamic capability networks [34], and CBAM networks, which amalgamate channel and spatial attention mechanisms [25].

Given the attention mechanism's role, it frequently serves as a module within neural networks to extract more representative and essential features through the fusion of channels and positional information. Consequently, this article constructs a spatial contextual attention module to enhance the robustness and transfer-ability of acquired information by introducing an attention mechanism.

C. Cross-Scene Object Recognition Migration

Remote sensing image scene understanding (RSISU) has emerged as a crucial task within the field of remote sensing, garnering attention from diverse research domains in recent years. In the context of RSISU, scene recognition within remote

sensing image analysis has also gained significant traction. Leveraging the widespread adoption of deep learning, the field of scene recognition has witnessed extensive integration of these techniques [35]. CNNs [36] have demonstrated remarkable achievements in natural image scene classification, and their application to remote sensing image scene classification holds great promise, bolstered by the migration of training models from datasets. Scene-oriented feature representation has been widely acknowledged as an effective strategy for interpreting high-resolution remote sensing images, with scene classification [27], [37], retrieval, and object detection holding substantial value for various critical applications [38]. However, scene recognition often grapples with the challenges of insufficient labeling. In this context, cross-scene migration provides a viable solution.

Cross-scene migration learning can be categorized into homogeneous migration learning and heterogeneous migration learning. When the target feature space and label space exhibit consistency, if $X_s = X_t$, $Y_s = Y_t$, then it is deemed homogeneous migration learning [39]. By adopting migration learning, we can address issues such as data scarcity and labeling limitations [40], computational constraints with abundant data, conflicts between ubiquitous models and individualized requirements, and the challenge of a cold start.

The pretrain+fine-tune approach, a classical migration learning model, belongs to the network migration-based category. While simplistic, it proves highly effective across numerous problems, sometimes rivaling domain adaptive and metalearning approaches. For example, Hossain et al. [41] proposed transfer learning fine-tuned on a deep convolutional neural network-based ResNet50 model to classify COVID-19 patients from a COVID-19 radiography database. However, relying solely on fine-tuning migration often leads to incomplete boundary and contextual information. To rectify this, scholars have proposed various strategies to enhance source domain feature extraction and bridge the gap between source and target domains. Zhu et al. [42] introduced the attention-based multiscale residual adaptation network to address cross-scene classification tasks. In a similar vein, Lu et al. [43] presented the multisource compensation network to tackle distribution discrepancies and incomplete categories within source and target domains in domain adaptation. In addition, an adversarial fine-grained adaptation network was developed to capture the intricate structures underlying data, improving discriminability and reducing domain disparities in class distributions [44]. Zhang et al. [45] developed a Single-source Domain Expansion Network to ensure the reliability and effectiveness of domain expansion. Tong et al. [46] proposed a channel attention-based DenseNet network for scene classification. This addresses the limitations of traditional stacked network structures in effectively extracting multiscale and pivotal features, thereby enhancing the overall feature representation capability.

III. Scenario Application and Problem Description

A. Scenario Application

The emergence of large-scale annotated data, the increase in computer computing power, and the development of advanced machine algorithms have all contributed to the prominence of deep learning. Within the field of deep learning, two primary methods are commonly utilized: 1) semantic segmentation; and 2) instance segmentation. Among these, semantic segmentation takes precedence, involving the classification of each pixel into distinct categories based on its value. Recognition of building areas in remote sensing images often grapples with challenges stemming from varying building scales and significant intraclass variability in buildings. Traditional CNNs exhibit limitations in comprehensively capturing overall building features due to their inherent structural constraints, making it challenging to grasp contextual dependencies between image patches. Consequently, this often results in subpar migration accuracy and undesirable model migration from source domain training to the target domain. To address this, a combined approach harnessing both Transformers and CNNs is employed. This amalgamation facilitates the extraction of global and local building features, enabling the recognition of building areas across different geographical regions and the identification of building areas in the target domain using limited samples through feature migration.

B. Problem Description

1) Building Classification: Deep learning can be described as a multistep procedure that leverages human mathematical insights and computer algorithms to construct a comprehensive framework. This architecture is then coupled with an extensive training dataset and the substantial computational capabilities of computers to iteratively fine-tune internal parameters towards achieving the desired problem objective. This process can be characterized as a semitheoretical and semiempirical modeling strategy. Suppose the input space is $\mathbf{T}$, the training dataset $\mathbf{T} = \{x_i, y_i\}_{i=1}^{N}$, where $x_i \in x$, $y_i \in y = \{0, 255\}$, $i = 1, 2, \ldots, N$; Prediction function See formula (1), loss function See formula (2)

$$y_{\text{pred}} = f(x, \theta), \tag{1}$$

$$Loss = -y \log y_{\text{pred}} - (1 - y) \log(1 - y_{\text{pred}}) \tag{2}$$

where y_{pred} is the predicted probability vector of the model. The goal of the training process is to find the optimal parameter θ so that the loss function $Loss$ reaches the minimum value.

2) Cross Scene Classification: Suppose a pretrained model with parameter θ_{pre}. This model has undergone training on the source task, which can be represented as shown in formula (3)

$$\theta_{\text{pre}*} = \arg\min_{\theta_{\text{pre}*}} (L_{\text{pre}}(\theta_{\text{pre}}, D_s)) \tag{3}$$

where $\theta_{\text{pre}*}$ is the optimal parameter of the pretrained model on the source task, and L_{pre} is the loss function of the source task. The training data for the source task is denoted as $D_s = (x_1, y_1), (x_2, y_2), (x_i, y_i), ..., (x_m, y_m), i = 1, 2, \ldots, m$, where x_i is an input sample for the source task, and y_i is the corresponding label.

The goal of the whole cross-scene migration problem is to find an optimal model parameter $\theta_{\text{pre}*}$ on the target scene, and to improve the learning performance of the target scene by minimizing the loss function $Loss$ on the data of the target scene,

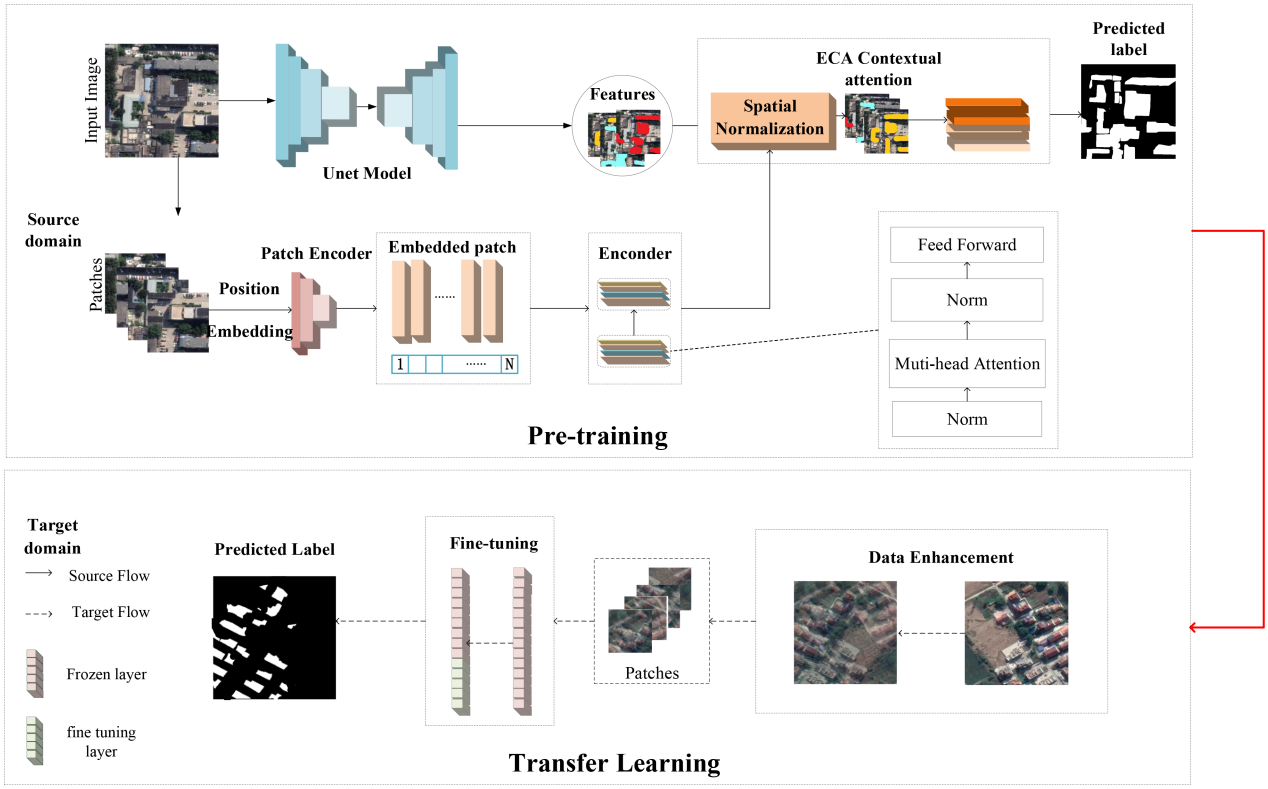

Fig. 2. Illustration of Combining attention mechanism and feature enhancement with Transformer and convolutional neural network model.

so that the knowledge of the source scene is effectively applied to the target scene.

The process of freezing plus fine-tuning in cross-transfer learning encompasses the following steps:

Step 1: Freezing Network Layers

First, certain layers within the pretrained model are frozen. Typically, this involves freezing the parameters of underlying or intermediate layers, preventing them from being updated during the subsequent fine-tuning phase. The frozen parameters can be represented as θ_{frozen}.

Step 2: Defining the fine-tuning model

Next, a fine-tuning model is defined based on the pretrained model. This fine-tuning model, represented by parameter $\theta_{\text{finetuned}}$, is designed to cater to the requirements of the target task. In addition, a fine-tuning loss function, denoted as $Lt(\theta_{\text{finetuned}}, D_t)$, is defined. This loss function is derived from the training data of the target task and the structure of the defined fine-tuning model.

Step 3: Conducting fine-tuning

During this step, the parameters of the fine-tuned model are updated. The objective is to minimize the fine-tuning loss function on the training data of the target task. This process can be mathematically expressed as in formula (4)

$$\theta_{\text{finetuned}*} = \arg\min_{\theta_{\text{finetuned}}}(Lt(\theta_{\text{finetuned}}, D_t)) \tag{4}$$

where $\theta_{\text{finetuned}*}$ is the optimal parameters of the fine-tuned model for the target task, the training data for the target task is denoted as $D_t = (x'_1, y'_1), (x'_2, y'_2), (x'_j, y'_j),..., (x'_n, y'_n), j = 1, 2, \ldots, n$, where x'_j is an input sample for the target task, and y'_j is the corresponding label.

IV. Methodology

A. Preliminary and Overview

In this section, we present the ECA-DSNN model, which incorporates both a dual-stream Network based on CNNs and Transformer, ECA-based feature fusion module, and a regularized data enhancement module into the decoder module. These additions aim to enhance the connectivity between building features. The model architecture is illustrated in Fig. 2. The essence of this approach can be summarized as follows.

1) Feature extraction based on dual-stream neural network: The method's core strategy involves complementing the local features derived from the CNN with the local feature information from the Transformer. This fusion amplifies the building recognition capabilities.
2) ECA Attention Mechanism: The utilization of the ECA serves to bolster interimage connection. This mechanism accentuates valuable image information while downplaying less relevant details. This integration contributes to achieving improved experimental outcomes, particularly with limited datasets.
3) Image regularization: Image regularization is employed to expand inherent image information. It proves adept at facilitating cross-domain migration within the target domain, even when dealing with constrained samples.

4) The proposed model further leverages dual parallel channels to augment building semantic features. By combining features extracted from both the Unet model and Transformer, the model's capability for recognizing buildings is further enhanced.

In addition, the spatial attention mechanism module is harnessed to capture attention across both channel and spatial dimensions. This module bolsters the model's ability to identify key features, and the data enhancement module is also incorporated to reinforce primary data features, effectively addressing issues of overfitting and subpar outcomes attributed to noise interference stemming from limited data. These enhancements collectively foster improved generalizability of the model.

B. Dual-Stream Feature Extraction Based on Dual-Stream Network Module

As depicted in Fig. 2, our model comprises two channels. In the upper channel, we execute pixel-level segmentation and modeling using the Unet structure. The given high-resolution image is characterized by three dimensions: $X \in R^{h\times w\times c}$, where h signifies the image height, w denotes the image width, and c denotes the image's channel count. In the lower channel, we initiate the process by chunking and flattening the image. For the input image, we divide the $h \times w \times c$ image into patches of uniform size, denoted as X_p, $X_p = [X_{1p}, X_{2p}, \ldots, X_{np}] \in R^{P^2\times c}$. Subsequently, each patch is fed into the transformer as a token, accompanied by positional embedding information denoted as $J_{\text{pos}} \in R^{N\times k}$, where k denotes the dimensionality of the space. The positional information, J_{pos}, is added to the embedding to establish patch order distinct, yielding $d_0 = [x1_{pj}; x2_{pj}; \ldots; xn_{pj}]$. Post this step, we apply dimensional scaling to the input data through the employment of the multihead attention (MSA) mechanism and MLP networks during patch embedding, feature learning, and classification. In the MSA layer, $d'_l = MSA(LN(d_l - 1)) + d_l - 1, l = 1, 2, \ldots, L$. Subsequently, in the MLP layer, $d_l = MSA(LN(d'_l)) + d'_l$, $l = 1, 2, \ldots, L$. Here, $LN()$ represents layer normalization, culminating in the categorization information, $y = LN(X^0_L)$.

Beyond feature encoding, we shape the image-level contextual representation (ICR) by reshaping the (Re) features, followed by a 1×1 convolution operation $(ConvI) : ICR = (Conv(Re(t_L)))$, yielding $ICR \in R^{h\times w\times 1}$. The ICR component is instrumental in constructing spatially dependent mappings at the image level, subsequently employed for normalization of the feature set generated by the CNN module.

C. ECA-Based Feature Fusion Module

To effectively extract building features and facilitate the aggregation of these extracted features, we implemented two key steps. First, we recalibrated the Unet building feature extraction module in the upper channel. Subsequently, we introduced spatial regularization to align the input features with the requirements for global feature extraction within the Transformer framework. In detail, we devised an attention mechanism module and introduced the channel feature parameter denoted as

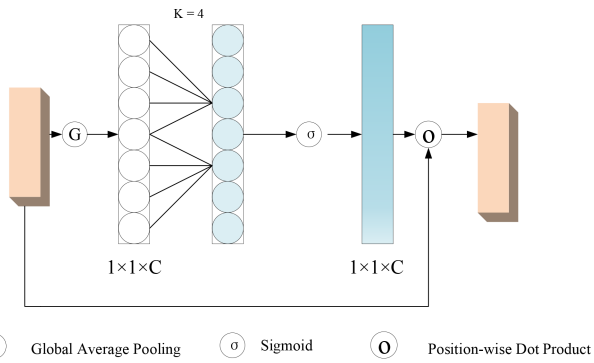

Fig. 3. Computational flow of W_{eca}.

W_{eca}. The calculation of W_{eca} is governed by the formula (5)

$$W_{\text{eca}} = Sigmoid(conv(AVG(leakyrelu(bn(conv(x)))))) \tag{5}$$

where x represents the input features meant for the global pooling by $b \times c \times h \times w$. These input features possess dimensions b, c, h, and w correspond to the dimensions of the input image. This operation transforms the input features into a sequence format of $b \times c \times 1 \times 1$. Through dimension adjustments and convolutional operations, this sequence is further moulded into a 1-D convolution pattern of $b \times 1 \times c$. A normalization step using Sigmoid is then applied to the resulting weights, facilitating dimensionality alignment to $b \times c \times 1 \times 1$. Ultimately, the obtained feature map is multiplied by the channel weights W_{eca}. The computational workflow is visually elucidated in Fig. 3.

We perform fusion between the output of the upper channel and that of the lower channel, thereby optimizing the features derived from the CNN using the outcomes of the lower channel. To incorporate spatial contextual information, we integrate the image-level ICR produced by the lower channel with the architectural features extracted by the upper channel. Before performing ECA dual-channel feature enhancement, we need to fuse the global feature T_{global} obtained from the dual-channel with the local feature U_{local} for feature fusion. This fusion serves to complement both spatial global and local information. By channeling the information from the upper channel Unet's extracted features into the spatial regularization, we enhance and supplement the information, yielding an output denoted as U_{local}. This U_{local} is then combined with the globally extracted spatial information T_{global} from the lower Transformer layer. This concatenation, symbolized as $X_{\text{out}} = cat[U_{\text{local}}, T_{\text{global}}]$, realizes the fusion process and generates the spliced output, which serves as the recognition results for ECA-based feature enhancement module's input.

D. Cross-Scenario Migration Module

Migration learning has proven effective in addressing limited labeled data in the target domain, enhancing convergence speed and training efficiency. To achieve this, we extend the model in the source domain using a combination of freezing and fine-tuning, coupled with regularization for feature enhancement. This approach aims to improve accuracy and expedite training

with a constrained sample size in the target domain. In the target domain dataset $D_t = (x'_1, y'_1), (x'_2, y'_2), \ldots, (x'_m, y'_m)$, we define the fine-tuned model, consolidating the parameters θ_{pre} and θ_{frozen} from the pretrained model into the unified parameter $\theta_{\text{finetuned}}$ for the fine-tuned model. We then establish the fine-tuning loss function $Lt(\theta_{\text{finetuned}}, D_t)$. To enhance the dataset, we leverage the Mixup method for data augmentation: $(x_{\text{aug}}, y_{\text{aug}}) = \text{mixup}(x, y, x', y')$. The exact calculation process can be represented by the Formula (6) and Formula (7)

$$x_{\text{aug}} = \lambda x + (1 - \lambda)x', \tag{6}$$

$$y_{\text{aug}} = \lambda y + (1 - \lambda)y'. \tag{7}$$

Here, (x, y) and (x', y') represent two randomly selected samples along their respective labels from the original training data. The resulting augmented samples and labels, denoted as $(x_{\text{aug}}, y_{\text{aug}})$ are generated through the MixUp process. Subsequently, we employ a suitable optimization algorithm, such as stochastic gradient descent (SGD). During the training phase, gradients are calculated, and the parameter $\theta_{\text{finetuned}}$ of the fine-tuned model is updated to minimize the fine-tuning loss function. This process is iterated multiple times until convergence is achieved or the predefined stopping condition is met.

V. Implementation

A. Datasets

Currently, commonly utilized public datasets for building extraction encompass the WHU building dataset, AIRS dataset, and Inria dataset. While these datasets serve as valuable resources, their foreign origins can lead to suboptimal results when employed for transfer learning to identify domestic buildings. To mitigate this, we leverage the Chinese typical urban building instance dataset [47] curated by Wu Kaishun et al., tailor-made for deep learning. Subsequently, the acquired model is subjected to transfer. Comprising 7260 region samples, this dataset is divided into three components: 1) images in.tif format; 2) .json format annotation files; and 3) .png format semantic segmentation labels. The image recognition model is trained using label compositions from.tif images and.png files. It encompasses four representative Chinese cities like Beijing, Shanghai, Wuhan, and Shenzhen. The dataset is standardized at a size of 500×500, subsequently cropped to 256×256 for the experiment. For our study, we have chosen to treat the Beijing dataset as the source domain, and a limited selection of Shanghai datasets as the target domain. Similarly, the Shanghai dataset is used as the source domain, while the Beijing dataset is employed as the target domain, thus facilitating our experimental analyses.

B. Evaluation Metrics

The experiment employs the following evaluation indicators: Accuracy, Precision, Recall, and F1 score. Each of these indicators is explained in detail below. In a binary classification scenario, instances are categorized into positive and negative classes, leading to four possible outcomes when classified by the classifier: True Positive (TP): Positive class correctly predicted as positive. False Negative (FN): Positive class incorrectly predicted as negative. False Positive (FP): Negative class incorrectly predicted as positive. True Negative (TN): Negative class correctly predicted as negative. These outcomes form the basis of the confusion matrix for evaluation metrics. Accuracy, Precision, Recall, and F1 are based on the above four situations predicted by the classifier

$$\text{Accuracy} = (\text{TP} + \text{FP})/(\text{TP} + \text{TN} + \text{FN} + \text{FP}), \tag{8}$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}), \tag{9}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}), \tag{10}$$

$$F1 = 2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall}). \tag{11}$$

C. Implementation Details

The deep learning framework used in this article is the Pytorch framework, which has good support for deep learning, so it is often used in machine learning and deep learning fields such as image recognition or speech recognition. The lab configuration of transferring is GPU: RTX3090, CPU 6 x Xeon Gold 6142, CUDA v11.2, Pytorch v1.10.

We randomly crop each image region into 256 × 256 size, effectively reducing the size of each image, increasing the speed of operation and facilitating subsequent processing. Our chosen optimizer is the Adam optimizer, which has a smoother gradient drop and can effectively prevent problems such as the SGD optimizer gradient plunge. The learning rate of the optimizer is 0.0001, and the training epoch is set to 50.

VI. Experimental Results

In this section, we merge the advanced transformers with CNNs to facilitate the identification of distinctive urban structures in China. By integrating various attention mechanisms and data enhancement modules, we introduce the ECA-DSNN model. Our analysis includes a comparison with existing models, further substantiating the excellence of the proposed approach in the field of building classification.

Baseline: For comparison, we selected several models including Unet, Transnet, Swin-Unet, and TMUnet, with TMUnet serving as the baseline. We evaluated both the original models and their augmented versions featuring attention mechanisms and mixup data augmentation, which is commonly employed in computer vision. This comparison aimed to confirm the enhanced effectiveness of our proposed mode.

The combination of Transformer and CNNs is predominantly used for feature classification in hyperspectral images within the remote sensing domain. However, its application to building classification in high-resolution images, particularly for domestic high-resolution images, remains limited. Thus, our initial focus is on validating the efficacy of integrating Transformer and CNN architectures for this purpose.

Pre-training: For the source domain model pretraining phase, we partitioned the Beijing dataset into training, test, and validation sets in a 7:2:1 ratio. We employed Unet, TMUnet, Transnet, and Swin-Unet for comparison purposes to assess the effectiveness and superiority of our proposed ECA-DSNN model. We conducted pretraining in the source domain using both the Beijing and Shanghai datasets. The outcomes for the

TABLE I
BEIJING SOURCE DOMAIN TEST SET

Category	Accuracy	Recall	Precision	F1
Unet	0.9443	0.7173	0.8429	0.7750
Transnet	0.9225	0.6284	0.7555	0.6861
Swin-Unet	0.9127	0.5622	0.7038	0.6251
TMUnet	0.9489	0.7453	**0.8538**	0.7959
ECA-DSNN	**0.9528**	**0.8222**	0.8266	**0.8244**

The bold values each of the columns represent the best among all of the models for each performance indicator.

TABLE II
SHANGHAI SOURCE DOMAIN TEST SET

Category	Accuracy	Recall	Precision	F1
Unet	**0.9429**	0.7644	0.7999	0.7817
Transnet	0.9152	0.6960	0.8192	0.7526
Swin-Unet	0.9034	0.7388	0.7468	0.7428
TMUnet	0.9125	0.7701	0.7668	0.7687
ECA-DSNN	0.9336	**0.8385**	**0.8207**	**0.8294**

The bold values each of the columns represent the best among all of the models for each performance indicator.

TABLE III
BEIJING SOURCE DOMAIN TEST SET TO SHANGHAI TARGET DOMAIN TEST SET

Category	Accuracy	Recall	Precision	F1
Unet	0.9592	0.7324	**0.7675**	0.7495
Transnet	0.9532	0.7414	0.7033	0.7218
Swin-Unet	0.9536	0.6675	0.7407	0.7023
TMUnet	0.9573	0.7475	0.7360	0.7417
ECA-DSNN	**0.9632**	**0.8282**	0.7632	**0.7864**

The bold values each of the columns represent the best among all of the models for each performance indicator.

TABLE IV
SHANGHAI SOURCE DOMAIN TESTSET TO BEIJING TARGET DOMAIN TESTSET

Category	Accuracy	Recall	Precision	F1
Unet	0.9222	0.6685	**0.7611**	0.7118
Transnet	0.8771	0.4983	0.5849	0.5381
Swin-Unet	0.9081	0.5909	0.7197	0.6490
TMUnet	0.8714	0.5926	0.5488	0.5699
ECA-DSNN	**0.9243**	**0.7045**	0.7531	**0.7279**

The bold values each of the columns represent the best among all of the models for each performance indicator.

test sets in both cities are presented in Tables II and III for reference.

Transfer learning: Utilizing the pretrained model Pk and Sh obtained from the aforementioned training, we conducted feature migration to the target domains of the Shanghai and Beijing datasets. The migration test outcomes were subsequently acquired and are presented in Tables IV and V for the Shanghai and Beijing datasets, respectively.

A. Pretraining

1) Beijing Source Domain: From Table I, it is evident that ECA-DSNN exhibits notable improvements compared to the benchmark TMUnet. Specifically, ECA-DSNN showcases a 0.39% enhancement in Accuracy, an impressive 7.69% improvement in Recall, and a noteworthy 2.85% boost in F1 score. Furthermore, compared with other commonly employed models, as well as CNNs combined with Transformers, ECA-DSNN emerges as the frontrunner in terms of both Accuracy and F1 score. These results effectively validate the efficacy of our ECA-DSNN model for building recognition within the Beijing dataset. A visual inspection of Fig. 4 underscores this advantage, as ECA-DSNN notably surpasses other models trained on the Beijing dataset, excelling in building integrity, contour edge clarity, and the recognition of block and strip buildings.

TABLE V
COMPLEXITY OF THE MODELS

Category	GFLOPs	Inference Time(ms/sample)	Params
Unet	62.238228	6.758	13,395,329
Transnet	64.457466	14.143	93,230,945
Swin-Unet	34.771610	7.694	41,341,536
TMUnet	196.636347	52.798	165,033,857
ECA-DSNN	196.636346	52.525	165,033,346

GFLOPs denotes gigafloat operations per second.

2) Shanghai Source Domain: Observing the results in Table II, in comparison to the Baseline TMUnet, ECA-DSNN demonstrates substantial improvements. Specifically, ECA-DSNN exhibits a 2.11% advancement in Accuracy, an impressive 5.39% boost in Precision, a significant 6.84% enhancement in Recall, and a notable 6.07% increase in F1 score. When compared to the Unet model, ECA-DSNN boasts a 4.77% higher F1 score and only a marginal decrease in accuracy by approximately 1%. These experiments effectively validate ECA-DSNN's efficacy across both the Beijing and Shanghai datasets, highlighting its adaptability and generalizability. As depicted in Fig. 5, concerning building zone identification, ECA-DSNN, TMUnet, and Unet all exhibit commendable results. However, upon closer examination, the Unet model presents a higher number of false positive and false negative predictions, while the TMUnet model's performance in predicting the contours of small-block buildings falls short of our proposed model. In terms of overall prediction efficacy, ECA-DSNN's predictions tend to align more closely with the true outcomes.

B. Cross-Scenario Migration

1) Beijing to Shanghai: Upon scrutinizing the accuracy of the migration from the Beijing dataset's source domain to the Shanghai dataset's target domain, as evaluated in Table III, ECA-DSNN demonstrates noticeable improvements. Specifically, ECA-DSNN showcases a 0.4% enhancement in Accuracy, a remarkable 9.4% surge in Recall, and an impressive 3.69% advancement in F1 score compared to the other comparative models. Furthermore, in contrast to the baseline model, ECA-DSNN's Accuracy registers a 0.59% rise, and its F1 score achieves a remarkable 4.47% boost. As illustrated in Fig. 6, it becomes evident that ECA-DSNN excels in the comprehensive recognition of building areas. Nevertheless, for contours and edges delineating small blocks of buildings, TMUnet encounters excessive misclassifications, failing to predict the edges' contours effectively. On the other hand, Unet exhibits a high incidence of misidentifications pertaining to small building blocks and displays subpar overall contour predictions. In a balanced evaluation, the predictions

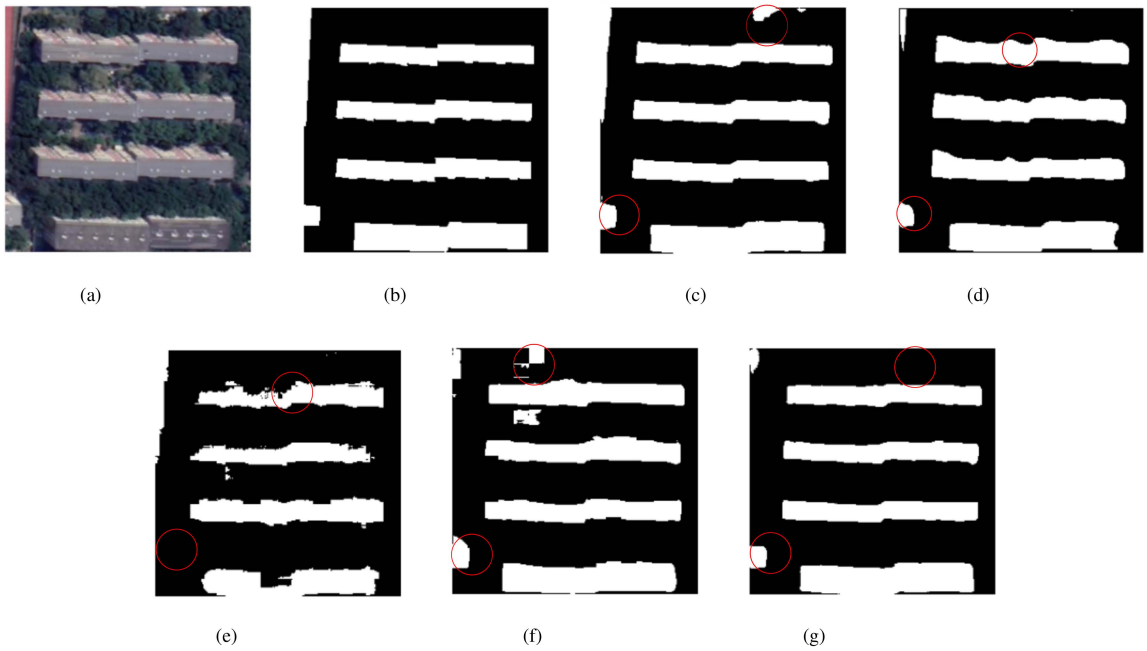

Fig. 4. Comparison of the real labels of Unet, Transnet, Swin-Unet, TMUnet, ECA-DSNN source domains with the prediction results in Beijing Source domain. (a) Sample. (b) Label. (c) Unet. (d) Transnet. (e) Swin-Unet. (f) TMUnet. (g) ECA-DSNN.

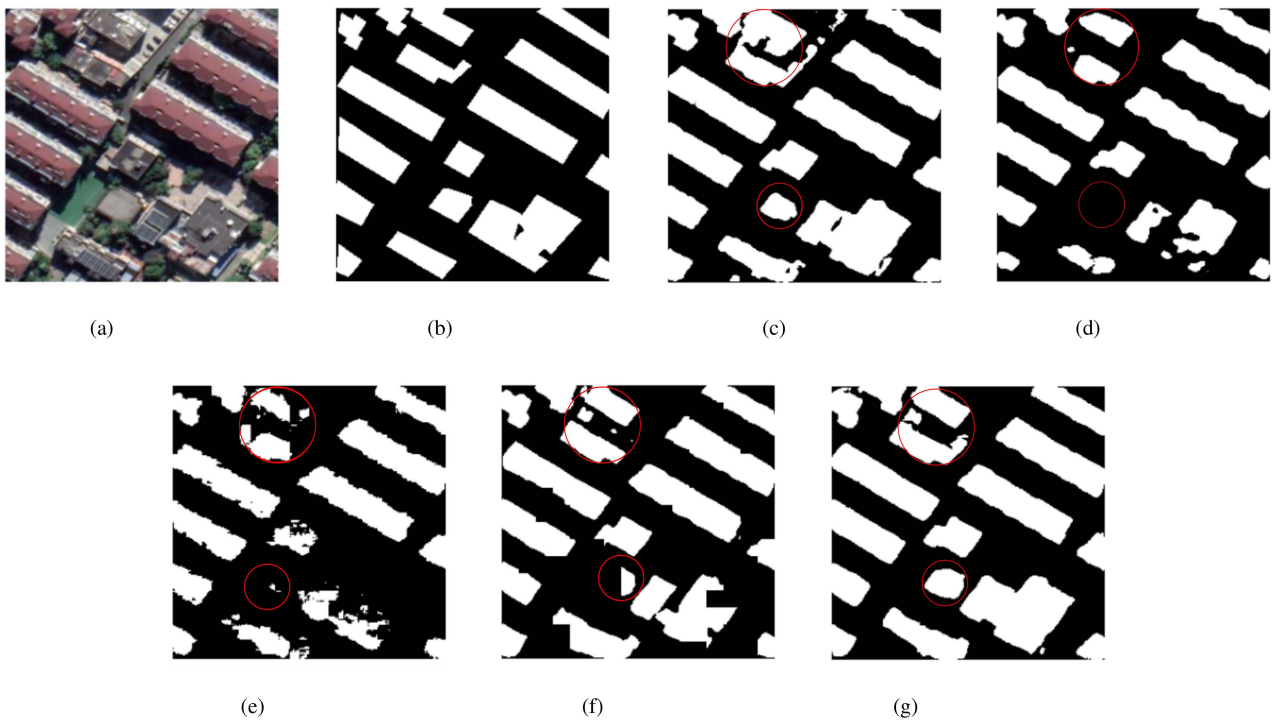

Fig. 5. Comparison of the real labels of Unet, Transnet, Swin-Unet, TMUnet, ECA-DSNN source domains with the prediction results in Shanghai source domain. (a) Sample. (b) Label. (c) Unet. (d) Transnet. (e) Swin-Unet. (f) TMUnet. (g) ECA-DSNN.

from ECA-DSNN tend to align more closely with the true outcomes.

2) Shanghai to Beijing: Examining the results of migrating from the Shanghai source domain dataset to the Beijing target domain dataset, as presented in Table IV, our innovative ECA-DSNN model demonstrates significant improvements. Specifically, ECA-DSNN showcases an impressive 5.29% increase in Accuracy and a remarkable 11.58% advancement in the F1 score compared to the benchmark model TMUnet. Among the comparative models, ECA-DSNN emerges as the most successful, yielding the finest outcomes. Fig. 7 vividly illustrates the outcomes across different models, revealing that ECA-DSNN,

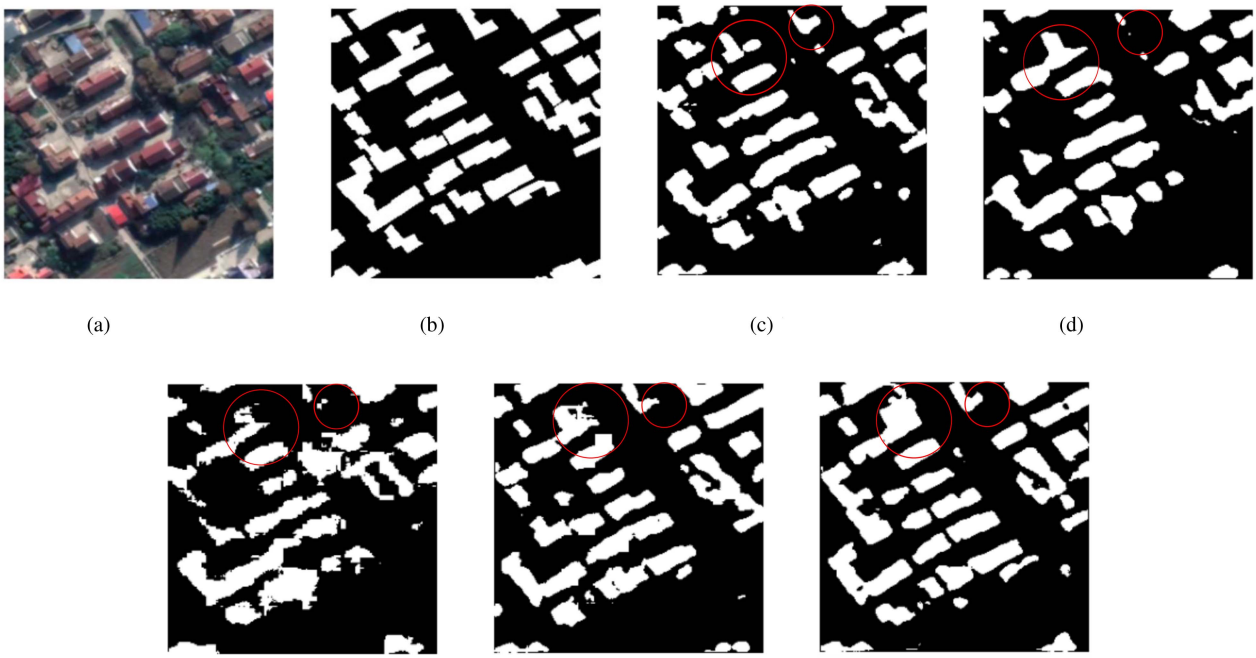

Fig. 6. Comparison of the real labels of Unet, Transnet, Swin-Unet, TMUnet, ECA-DSNN in Shanghai target domain and Beijing source domain with the prediction results. (a) Sample. (b) Label. (c) Unet. (d) Transnet. (e) Swin-Unet. (f) TMUnet. (g) ECA-DSNN.

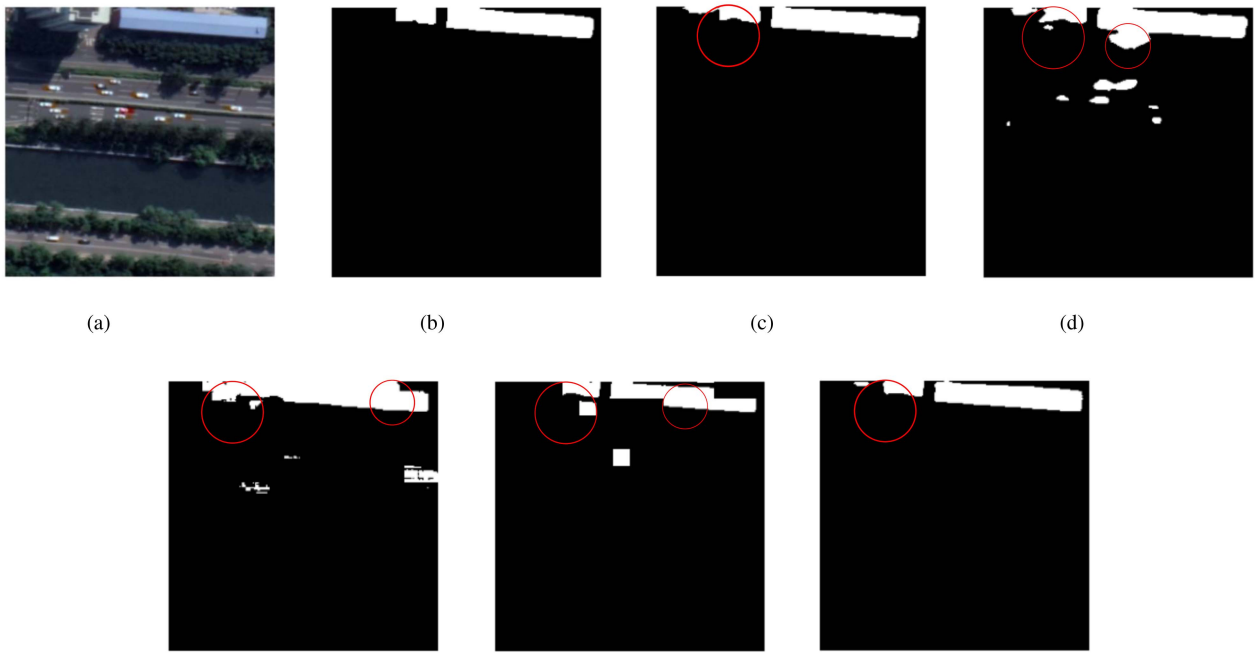

Fig. 7. Comparison of the real labels of Unet, Transnet, Swin-Unet, TMUnet, ECA-DSNN in Beijing target domain and Shanghai source domain with the prediction results. (a) Sample. (b) Label. (c) Unet. (d) Transnet. (e) Swin-Unet. (f) TMUnet. (g) ECA-DSNN.

TABLE VI
IMPACT OF THE DIFFERENT MODULES ADDED ON THE PROPOSED MODEL

Transformer	Mixup	ECA	Accuracy	Recall	Precision	F1
×	×	×	0.9592	0.7324	0.7675	0.7495
✓	×	×	0.9578	0.7052	0.7341	0.7193
✓	✓	×	0.9549	0.7475	0.7360	0.7417
✓	×	✓	0.9629	0.7476	0.7360	0.7712
✓	✓	✓	**0.9632**	**0.8282**	0.7632	**0.7864**

The bold values each of the columns represent the best among all of the models for each performance indicator.

fortified by the incorporation of the data enhancement module and the contextual relationship module, emerges as the top performer.

Furthermore, we undertook a calculation to determine the computational complexity of the theoretical model and compared it with the aforementioned alternative methods, as exemplified in Table V. FLoating point OPerations per second (FLOPs) represents the total number of accumulations of the model's floating-point calculations. It can be used to measure the complexity of the model. Model inference time is the time it takes for the model to process a picture or a sample.

From the table, it can be found that the FLOPs of ECA-DSNN is greater compared to Unet, Transnet, Swin-Unet, and lower than that of TMUnet. It is also found that the inference time of the model is not only related to FLOPs, but also to the number of parameters. For instance, although the complexity of Unet and Transnet models are similar, the number of parameters of the Unet is much smaller than that of Transnet, and thus the inference time of the model Unet is also lower than Transnet. Moreover, in terms of inference time, the proposed method takes longer time for theoretical reasoning compared to Unet, Transnet, and Swin-Unet. Although the FLOPs of the ECA-DSNN are not much different from those of the TMUnet, the model parameters of the ECA-DSNN are slightly smaller than those of the TMUnet, and thus the inference time of the proposed model is a little bit faster.

C. Ablation Experiments

The additional network incorporates the attention mechanism module and the data enhancement module. We chose to remove each of these modules to assess the impact of each module on the model's performance improvement. As shown in Table VI, we can observe that the removal of any of these modules has a noticeable impact on the model's performance. The elimination of the ECA-based feature fusion module significantly reduces the model's contextualization capability, making it more susceptible to noise and negatively impacting the learning of negative example samples. Consequently, this leads to a reduction in performance and an increase in instability. On the other hand, the omission of the Mixup module results in the model learning the key features of the dataset with less prominence. This weakens the model's convolutional mapping ability, ultimately causing a decrease in performance.

The accuracy improved by 0.37% and F1 score improved by 2.37% when compared to the results before and after not adding the data enhancement module. In addition, the accuracy improved by 0.83% and F1 score improved by 4.47% when compared to the results before and after adding the attention mechanism module. These comparisons underscore the utility of the different modules in enhancing model performance and highlight their essential role.

VII. CONCLUSION

In this article, a novel dual-stream feature extraction module, ECA-based feature fusion module and a regularization module are introduced, which are integrated into the benchmark model TMUnet, and applied to the problem of finite-sample cross-scene migration. Upon these, the ECA-DSNN model is proposed. These added modules in the model aim to address the shortcomings of the model's lack of spatial context learning capability and insufficient shallow data feature learning, thereby significantly improving the accuracy of the pretrained model.

Compared to the benchmark model, the proposed model shows excellent generalization ability when migrating across different data domains. The trained model proved to be more adaptable to different data domains. Comparing different models with ECA-DSNN in terms of learning building detail features and recognizing building edges. The proposed model demonstrates proficient performance in recognizing detail features and building edge contours. This achievement can be attributed to the combination of contextual spatial relationships and sample feature enhancement.

However, there are still specific limitations with our method. Due to the relatively enlarged model parameters, the duration for training is extended, consequently leading to higher costs compared to some simpler deep learning models with fewer layers. Despite the higher accuracy achievable with similar computing power, shortening the training time remains a challenge. Furthermore, although the model achieves relatively excellent accuracy with a limited number of samples, a relatively substantial amount of source domain data is still required. Therefore, training a more generalized model with limited source domain data remains a future consideration.

REFERENCES

[1] L. Deren, Z. Liangpei, and X. Guisong, "Automatic analysis and mining of remote sensing Big Data," *Acta Geodaetica et Cartographica Sinica*, vol. 43, no. 12, 2014, Art. no. 1211.
[2] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1662.
[3] S.-Y. Ji and H.-J. Jun, "Deep learning model for form recognition and structural member classification of east asian traditional buildings," *Sustainability*, vol. 12, no. 13, 2020, Art. no. 5292.
[4] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, no. 2, pp. 1097–1105, 2012.
[5] Z. Yang, H. Yu, Y. He, W. Sun, Z.-H. Mao, and A. Mian, "Fully convolutional network-based self-supervised learning for semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 132–142, Jan. 2024.
[6] W. Deng, Q. Shi, and J. Li, "Attention-gate-based encoder–decoder network for automatical building extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2611–2620, 2021.
[7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represenattions*, 2021.

[8] L. Chen et al., "Decision transformer: Reinforcement learning via sequence modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 15084–15097.
[9] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[10] X. Zhang et al., "A-GCRNN: Attention graph convolution recurrent neural network for multi-band spectrum prediction," *IEEE Trans. Veh. Technol.*, vol. 73, no. 2, pp. 2978–2982, Feb. 2024.
[11] R. Azad, M. Heidari, Y. Wu, and D. Merhof, "Contextual attention network: Transformer meets U-net," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2022, pp. 377–386.
[12] D. Guo, Y. Xia, and X. Luo, "Scene classification of remote sensing images based on saliency dual attention residual network," *IEEE Access*, vol. 8, pp. 6344–6357, 2020.
[13] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.
[14] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[15] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, pp. 1–40, 2016.
[16] Y. Luo, T. Liu, D. Tao, and C. Xu, "Decomposition-based transfer distance metric learning for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3789–3801, Sep. 2014.
[17] S. Moon and J. G. Carbonell, "Completely heterogeneous transfer learning with attention-what and what not to transfer," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 1–2.
[18] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.
[19] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," 2018, *arXiv:1803.02155*.
[20] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 15908–15919.
[21] L. Yuan et al., "Tokens-to-token ViT: Training vision transformers from scratch on imagenet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 558–567.
[22] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1245–1256, Jun. 2017.
[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[24] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3156–3164.
[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[26] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "PGA-siamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 484.
[27] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, Apr. 2015.
[28] M. F. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber, "Deep networks with internal selective attention through feedback connections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, vol. 2, pp. 3545–3553.
[29] L. Wang, J. Peng, and W. Sun, "Spatial–spectral squeeze-and-excitation residual network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 884.
[30] Y. GE, L. MA, F. YE, and J. CHU, "Remote sensing image retrieval based on multi-scale pooling and norm attention mechanism," *J. Electron. Inf. Technol.*, vol. 44, no. 2, pp. 543–551, 2022.
[31] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*.
[32] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11534–11542.
[33] M. Jaderberg et al., "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 28, pp. 2017–2025.
[34] A. Almahairi, N. Ballas, T. Cooijmans, Y. Zheng, H. Larochelle, and A. Courville, "Dynamic capacity networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2549–2558.
[35] Y. Gu, Y. Wang, and Y. Li, "A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection," *Appl. Sci.*, vol. 9, no. 10, 2019, Art. no. 2110.
[36] O. A. Penatti, K. Nogueira, and J. A. Dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit. workshops*, 2015, pp. 44–51.
[37] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. geographic Inf. Syst.*, 2010, pp. 270–279.
[38] X. Zhang and S. Du, "A linear dirichlet mixture model for decomposing scenes: Application to analyzing urban functional zonings," *Remote Sens. Environ.*, vol. 169, pp. 37–49, 2015.
[39] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.
[40] Y. Luo, Y. Wen, L.-Y. Duan, and D. Tao, "Transfer metric learning: Algorithms, applications and outlooks," 2018, *arXiv:1810.03944*.
[41] M. B. Hossain, S. H. S. Iqbal, M. M. Islam, M. N. Akhtar, and I. H. Sarker, "Transfer learning with fine-tuned deep cnn resnet50 model for classifying covid-19 from chest x-ray images," *Inform. Med. Unlocked*, vol. 30, 2022, Art. no. 100916.
[42] S. Zhu, B. Du, L. Zhang, and X. Li, "Attention-based multiscale residual adaptation network for cross-scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5400715.
[43] X. Lu, T. Gong, and X. Zheng, "Multisource compensation network for remote sensing cross-domain scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2504–2515, Apr. 2020.
[44] S. Zhu, F. Luo, B. Du, and L. Zhang, "Adversarial fine-grained adaptation network for cross-scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2369–2372.
[45] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, "Single-source domain expansion network for cross-scene hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1498–1512, 2023.
[46] W. Tong, W. Chen, W. Han, X. Li, and L. Wang, "Channel-attention-based densenet network for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4121–4132, 2020.
[47] K. Wu et al., "A dataset of building instances of typical cities in China," *Chin. Sci. Data*, vol. 6, no. 1, pp. 182–190, 2021.

Wenmei Li (Member, IEEE) received the M.S. degree in cartography and geographic information systems from Nanjing University, Nanjing, China, in 2010, and the Ph.D. degree in forest management-remote sensing technology and application direction from the Chinese Academy of Forestry, Beijing, China, in 2013.

She is currently a Professor with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China. Her research interests include deep learning, optimization, 3-D reconstruction, and their utilization in the field of land remote sensing.

Jiadong Zhang received the B.S. degree, in 2022, in geographic information science from the Nanjing University of Posts and Telecommunications, Nanjing, China, where he is currently working toward the M.A. degree in surveying and mapping science and technology.

His research interests include deep learning and inversion of forest parameters.

Hao Xia received the B.S. degree, in 2023, in human geography and urban rural planning from the Nanjing University of Posts and Telecommunications, Nanjing, China, where he is currently working toward the M.A. degree in surveying and mapping science and technology.

His research interests include PolSAR image processing and deep learning.

Qing Liu received the B.E. degree in remote sensing science and technology from Chang'an University, Xi'an, China, in 2021. He is currently working toward the M.A. degree in surveying and mapping science and technology with the Nanjing University of Posts and Telecommunications, Nanjing, China.

His research interests include deep learning and its applications in remote sensing image processing.

Yu Wang (Member, IEEE) received the Ph.D. degree in signal and information processing from the Nanjing University of Posts and Telecommunications (NJUPT), Nanjing, China, in 2023.

Since 2023, he has been an Appointed Professor with the Nanjing University of Posts and Telecommunications, Nanjing, China. He has authored or coauthored more than 15 papers in peer-reviewed IEEE journal/conferences. His research interests include deep learning, optimization, and its application in wireless communications.

Dr. Wang was the recipient of three Best Paper Awards–CSPS 2018, CSPS 2019, and ICEICT 2019.

Yan Jia received the double M.S. degree in telecommunications engineering and computer application technology from Politecnico di Torino, Turin, Italy, and Henan Polytchnic University, Jiaozuo, China, in 2013, and the Ph.D. degree in electronics engineering from Politecnico di Torino in 2017.

She is currently with the Nanjing University of Posts and Telecommunications, Nanjing, China. Her research interests include microwave remote sensing, soil moisture retrieval, GNSS-R applications on land remote sensing, and antenna design.

Yixiang Chen received the Ph.D. degree in cartography and geographic information engineering from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2013.

He is currently with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include remote sensing image processing, geospatial information extraction, and urban remote sensing.