Heritage Science

RESEARCH ARTICLE

Open Access

Exploring spatiotemporal changes in cities and villages through remote sensing using multibranch networks

Zhao Mengqi[1*] and Tian Yan[1,2]

Abstract

With the rapid development of the social economy, monumental changes have taken place in the urban and rural environments. Urban and rural areas play a vital role in the interactions between humans and society. Traditional machine learning methods are used to perceive the massive changes in the urban and rural areas, though it is easy to overlook the detailed information about the changes made to the intentional target. As a result, the perception accuracy needs to be improved. Therefore, based on a deep neural network, this paper proposes a method to perceive the spatiotemporal changes in urban and rural intentional connotations through the perspective of remote sensing. The framework first uses multibranch DenseNet to model the multiscale spatiotemporal information of the intentional target and realizes the interaction of high-level semantics and low-level details in the physical appearance. Second, a multibranch and cross-channel attention module is designed to refine and converge multilevel and multiscale temporal and spatial semantics to perceive the subtle changes in the urban and rural intentional targets through the semantics and physical appearance. Finally, the experimental results show that the multibranch perception framework proposed in this paper has the best performance on the two baseline datasets A and B, and its F-Score values are 88.04% and 53.72%, respectively.

Keywords: Intentional target, Spatiotemporal changes, Multiscale spatiotemporal, Cross-channel attention, Information interaction

Introduction

With the continuous development of the social economy, human living standards have undergone tremendous changes. Cities and villages, as gathering places for human social interaction and activities, have also experienced massive changes in recent years. In addition, the cities and villages that people call home not only reflect their lifestyles, but also affect their physical health, mental health and social well-being. Exploring the urban and rural environmental changes from the acquired remote sensing data helps to understand the development of society and the economy in depth. At the same time, it can also effectively judge whether it is necessary to further improve the infrastructure construction and the quality of life in these urban and rural spaces.

In recent years, with the application of computer intelligence interpretation technology in many fields, such as natural language processing (NLP), image classification, and object detection (OD), it has provided a new way of evaluating the city and village environmental changes (including buildings, infrastructure and heritage). In the early stages of urban and rural change research, people usually use a variety of different methods to simulate and measure the construction, cultural heritage, infrastructure, and environment of a certain area, by using digital models to obtain useful information and build urban forms and urban environments. However, with

*Correspondence: zhaomengqi@whut.edu.cn
[1] School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan 430000, China
Full list of author information is available at the end of the article

the complexity of urban and rural environments, digital modeling also has difficulty meeting the increasing application requirements. At the same time, affected by the exponential growth of big data, it is difficult to obtain effective detailed information with this kind of simulation modeling method. Meanwhile, digital modeling is usually oversimplified and therefore, is unavailable not for some studies, including studies of the changes in infrastructure. Additionally, because it neglects the netural landscapes of cities and villages, it has proven to be less effective.

However, to obtain more detailed information from relevant data to effectively simulate urban and rural environmental changes, many machine learning and deep learning algorithms have been developed. For instance, Naik et al. [1] to rate the safety, resident wealth and vitality index of the block, input the data collected from Google into the machine learning model for modeling, and generate new neighborhood semantic information. Gebru et al. [2] proposed a deep learning method to estimate different choices in the United States, the socioeconomic situation of the district, and the methods using a large number of geotagged street images. Li et al. [3] presented urban landscape study methods by combining deep convolutional neural networks (DCNNs) and street-level images, which accurately recognized the different urban features from these street-level images. Meanwhile, machine learning and deep learning methods also have strong modeling ability for complex and large-scale data, applying these methods to large-scale urban complex data, such as occlusion and zoom, and learn the location or category of the target object through supervised training [4, 5]. Obeso et al. [6] adopted deep convolutional natural network methods to train and predict visual attention in natural images to address the classification problem of Mexican cultural heritage. Morbidoni et al. [7] proposed novel methods for learning from synthetic point cloud data for historical building semantic segmentation, mainly to provide a first assessment of the use of synthetic data to drive convolutional network-based semantic segmentation in the context of historical buildings.

Although the above methods detect urban and rural environments to a certain extent, they basically focus on segmentation tasks, such as content classification and recognition of buildings, while ignoring the changes in the urban and rural spatiotemporal environment. At the same time, these methods ignore the image feature extraction process. There are subtle changes in the target object and a poor perception of the temporal and spatial semantics. Thus, we address these issues and explore changes in the urban and rural environment, as well as the form and infrastructure from the perspective of time and space. We present a novel spatiotemporal perception method to explore changes in cities and villages from remote sensing with multibranch networks. We aim to build visual spatiotemporal perception models that can be used to estimate environmental, form and infrastructure changes in urban areas and villages, while vigorously promoting the development of social research and improve the lifestyle of humans.

In summary, the main work in the paper is as follows:

- Frameworks: A methodology for exploring the spatiotemporal changes from remote sensing of city and village environments, forms and infrastructure aspects. The main aim is to build a relationship between human visual perspectives and perceptions that can understand the changes in social development, and improving the effectiveness of statistics from society.
- Technology: We present a novelty perception frameworks using multibranch networks. This method mainly uses a multibranch attention network to model remote sensing images in the same area at different time periods, forming information sharing in time and space. Second, through this information, the model can perceive subtle changes in different targets in cities and villages, including target positions, physical structures and geometric shapes. It further establishes temporal and spatial dependence on different scales to generate better representations to complete relevant statistics and reasoning.
- Application: For the application of subsequent tasks, such as urban planning, intention target statistics, disaster evaluation, etc. based on baseline datasets a and b, using preprocessing methods with rotation and noise addition, the perception framework proposed in this paper is tested and verified. The final experimental results show that our proposed framework has achieved good experimental results and perceives the average area of urban and rural intentional changes.

The rest of the organizational structure of the paper is as follows: In Section 2, we elaborate on the related work of urban and the image perception of village environments, forms, infrastructure, etc. Section 3 describes our proposed perception frameworks in detail. Section 4 discusses and analyzes the processing of datasets and the application. Then, we present the detailed experimental results and describe the changes. Section 5 provides a brief summary and possibilities for future work.

Related works

In this section, we elaborate on the related work on the image perception of form, infrastructure, etc. in cities and villages. The primary is divided into traditional and deep neural networks of urban and village environmental forms or the infrastructure's architectural elements in visual content. It is worth noting that deep neural network methods mainly focus on tasks, such as image classification, segmentation and detection.

The traditional image perception of cities and villages

Currently, image datasets have been widely used in many files of urban and village research and planning in progress; for example, the main application files contain regional city systems, city and village spatial structures, infrastructure service systems, transportation and travel and collective activities in society. However, with the continuous development of society and the economy, people's application needs are gradually increasing. It is time-consuming and expensive to use manual statistics to collect relevant information. Thus, many researchers have developed algorithms to perceive urban and rural areas from different perspectives and archive the effective information, such as the form and environment of the plant. For instance, Hu et al. [8] proposed an effective method of typology analysis, which entailed they using computer technology to check the content of different images and adopting clustering methods to judge the activity levels of the different types of users on Instagram. Hochman et al. [9] proposed a method based on Instagram algorithms, which was a spatiotemporal pattern analysis method designed to visualize the characteristics of image content from 13 different cities around the world and make corresponding comparisons to further describe people's daily activities, culture, etc. However, to facilitate the interaction of users and existing image datasets and further extend the scale of these image datasets via social media, Jett et al. [10] present a feedback framework for transferring user-generated information to institutional data providers, which can improve the service scope of the dataset center. However, the methods mainly use cultural heritage institutions that can also enhance collections by sharing content through popular web services. The abovementioned methods mainly use some simple visual methods to analyze the images of cultural heritage, residents' living conditions and their environments circulating on social media during the disaster. Although quick and simple statistics are realized to further expand the relevant database, it is not possible to perceive changes from a deeper level, such as damage to residential areas, cultural buildings and other infrastructure in the disaster.

However, there are also many researchers who focus on identifying urban or rural building structures from natural images generated by users and analyzing the relevant characteristics of buildings. For example, Li et al. [11] addressed the sustainable development problem of cities and the effective identification of urban functional areas. They combined multisource geographic data to establish a quantitative measurement method for urban functional areas. Bose et al. [12] take the Siliguri metropolitan area in West Bengal, India as the research object, propose novelty study methods of the Markov chain model and analyze the spatial distribution of urban land. Liuet al. [13] scientifically plan the urbanization layout and improve the utilization rate of land space. Urban functional areas are identified and analyzed from the perspective of data mining, and taxi trajectory data are used as the research basis for urban functional areas. A DTW-based approach is proposed. K-nearest's classification algorithm for cluster recognition of urban functional areas. Although these methods can effectively identify the functional areas of the city, they have not effectively combined the temporal and spatial information of the city and the countryside in the analysis and statistics process. When the environment is complex, it is difficult to distinguish the functional areas efficiently and accurately. The cultural heritage, buildings and roads in the functional area are not analyzed in detail.

Conversely, many researchers pay more attention to the perception of the form and infrastructure of residential areas in urban functional areas, such as Tardioli, Giovanni and Kerrigan, Ruth et al. [14], to evaluate the building energy in the city. A new method is proposed to identify building clusters, and a dataset of representative buildings is provided. At the same time, the method is mainly divided into three parts: building classification, building clustering and prediction. Gadal, S Bastien and Ouerghemmi, Walid et al. [15] considered that hyperspectral remote sensing images can describe surface objects and landscapes more accurately, and a classification method based on an urban target spectral database was proposed to detect and classify specific urban targets. Manzoni, Marco and Monti-Guarnieri, Andrea et al. [16] combined synthetic aperture radar (SAR) images and geospatial information systems, developing a simple and fast method to identify structural changes in buildings in urban environments. This proposed method can effectively evaluate small changes after disasters.

The deep learning image perception of cities and villages

Although these methods can reduce the errors caused by hand-made features, in a complex environment, it is difficult to effectively capture the detailed changes of the target (such as buildings, roads, bridges, etc.) in the form, physical structure, or geometric form in the image using simple machine learning. Thus, deep learning techniques are widely used in tasks, such as urban planning, urban building classification, and urban form perception. Llamas, Jose and M Lerones, Pedro et al. [17] present a novel method of the classification of architectural heritage images with deep convolutional neural networks.

The main objective of this article is to introduce the application of techniques based on deep learning for the classification of images of architectural heritage, specifically through the use of convolutional neural networks. Meanwhile, the methods can achieve better management and a more effective search of the urban architectural heritage. They are also beneficial for the tasks of studying and interpreting the heritage asset in question. With the rapid development of urban areas and villages, due to their wide distribution, construction waste is easily confused with the surrounding environment and difficult to manually classify. At the same time, traditional single-spectral feature analysis has difficulty extracting and identifying urban construction waste-related information. Thus, Chen et al. [18], utilizing the multifeature analysis method of remote sensing images, developed a method for extracting urban construction waste information from the optimal VHR image combined with a morphological index and hierarchical segmentation. Attari et al. [19] assessed the extent of damage to urban and village building structures after the disaster, and with UAV imagery, proposed a fine-grained classification method called Nazr convolution neural networks (Nazr-CNN) to conduct a damage assessment. Vetrivel et al. [20] suggested that to improve the performance of damage detection, the CNN and 3D point cloud information of the target object in the image are, respectively extracted, and the multicore learning framework is used to combine the two kinds of information to achieve classification, while finally performing damage detection on the building roof and other object. Subsequently, Hamdi et al. [21] presented a forest damage assessment method with deep learning techniques, and the backbone network of the method was mainly U-Net. Although these methods have achieved good results in the postdisaster assessment, they mainly focus on the use of UAV images and hyperspectral remote sensing images.

In recent years, some researchers have used images collected on social media to perceive the ideology of cities and villages. For example, in the case of disasters and a lack of labeled data, Li et al [22] proposed a domain-adaptive countermeasure neural network method to recognize disaster images and detect damaged areas. Meng et al. [23] verified the correlation between the physical health of the elderly and the urban space using the Baidu Street View (BSV) of the Macau Peninsula as the research scene, and deep learning technology was used to perceive the high-density urban street space. Kim et al [24] proposed understanding tourists' urban images with geotagged photos using convolutional neural networks. With the continuous increase in the urban population, the human gathering area has gradually evolved into a local dense temporal and spatial dynamic distribution. To better understand the urban environment, Chen et al [25] constructed an advanced image recognition model and used marked Flickr pictures to train the neural network to quantify the feature information of different cities. Jayasuriya et al [26] presented a novel localizing PMD perception method for urban streets via convolutional neural networks. The method combines two important components, one of which uses a CNN to extract the feature information of infrastructure such as roads, lane markings, and manhole covers and form a location. The other component is mainly to use a CNN to detect common environmental landmarks, such as tree trunks for positioning. However, to further enhance a human's perceptibility for urban and village forms, the environment and the infrastructure, Wang et al. [27] presented a new multitask and multimodal deep learning framework with automatic loss weighting to assess the damage after disastrous events. Agarwal et al [28] proposed multimodal damage analysis methods to reply to deployment, challenges and assessment and are called Crisis-DIAS. In addition, other related two-branch neural networks, such as the Fractional Gabor Convolutional Network (FGCN) was proposed by Zhao et al. [29, 30]. The information fusion and Patch-to-Patch CNN uses remote sensing image tasks by Zhang et al. [31, 32] with the manner of word embedding using image processing [33].

In summary, although the above methods use deep learning technology to improve people's perceptions of the social environment and form, most of them use simple deep learning methods to classify, segment, and detect corresponding image data, which are not sensitive to spatiotemporal information. At the same time, in the process of target feature extraction, a large amount of detailed information is ignored, which makes the feature information unable to effectively describe the target (urban and rural buildings, roads, etc.), ultimately leading to large perceptual errors. Second, these methods do not take into account the changes in the same area at different time periods.

Our proposed methods

In this section, we will elaborate on our proposed spatiotemporal perception framework from three aspects: the feature extraction of urban and village images, the network structure of the backbone and the adjustment and optimization.

Overview

With the rapid development of society and the economy and field surveys of urban and rural residents' gathering places or other nongathering places, it can be found that there are huge differences in the forms, environments, and infrastructures presented in different regions and at different times. For example, the distribution of residential areas and functional areas is irregular. At the same time, the distribution of the environment and infrastructure also changes with changes in the gathering place. However, when external factors are more complex, if using traditional machine learning methods to perceive changes in the same area at different periods of time, people are susceptible to interference from these external factors because of light and occlusion, resulting in larger perception errors and affecting subsequent applications. The deep learning method has a strong self-learning ability, and can use the activation state of the neurons in the network structure to capture the detailed information of the urban and rural targets in the image, as well as high-level abstract distinguishable information to improve the perception accuracy. Therefore, subtle changes in the urban and village environment, form and infrastructure in different time periods are detected from the limited remote sensing data to improve the perception accuracy and the efficiency of subsequent applications, such as the statistics of urban planning and environmental information. We propose a spatial-temporal sensing method to detect urban and rural changes from the perspective of remote sensing. The method mainly includes spatial branches and temporal branches. The temporal branch embeds the urban and village images in the same area in different time phases to enhance the interaction between images in different time phases and establish effective dependencies. For spatial branching, the main purpose is to model the target object in the image to form a strong difference within or between classes so that it has better recognition. The network structure of our proposed spatiotemporal perception framework is shown in Fig. 1.

Considering that urban and rural images in the same area at different times have both relevance and spatial and temporal differences, we set the input images to

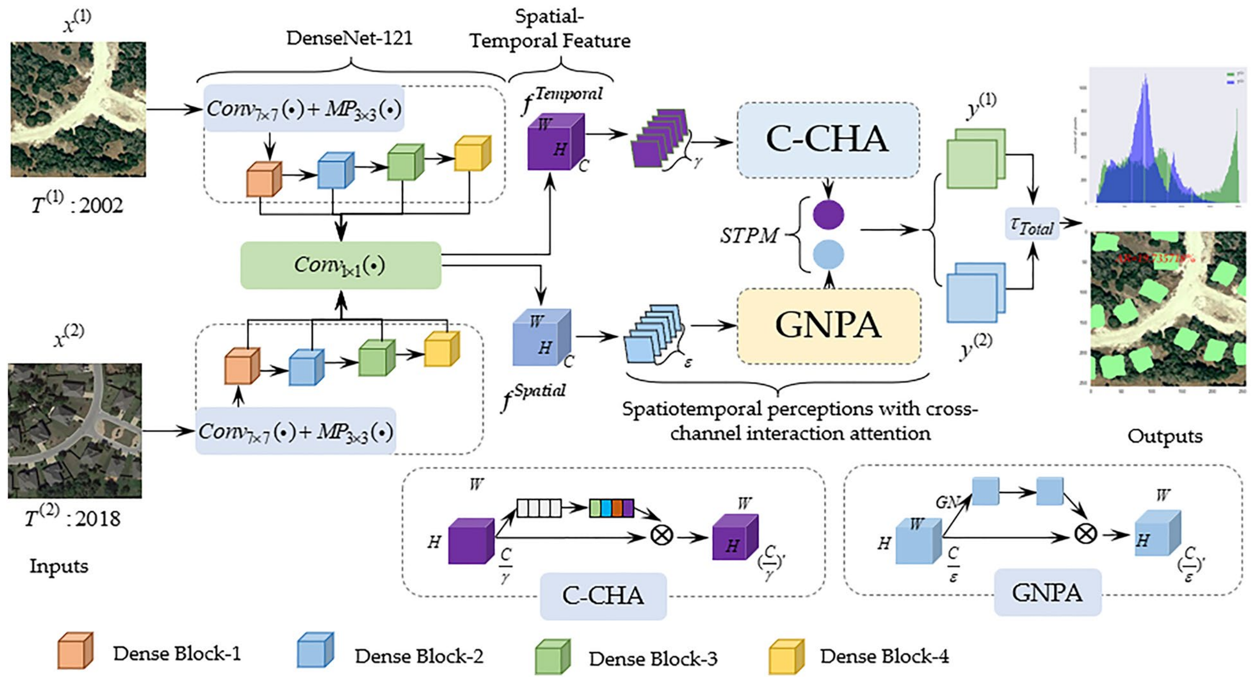

Fig. 1 The network structure of our proposed perception frameworks. where $T^{(1)}:2002$ and $T^{(2)}:2018$ indicate different time phases. $x^{(1)}$ and $x^{(2)}$ indicate the remote sensing images of the input. $f^{Spatial}$ and $f^{Temporal}$ indicate the spatial information and temporal information via the feature extraction module, and the module mainly contains densely connected convolutional networks (DenseNet-121). *H*, *W*, *C* indicates the height, width and channel, respectively. γ, ε indicates the subspace of temporal and spatial feature maps, $(\frac{C}{\gamma})\prime = \frac{C}{8\gamma}\prime(\frac{C}{\varepsilon})\prime = \frac{C}{8\varepsilon}.y^{(1)}$ and $y^{(2)}$ indicates the output feature via STPM, where STPM indicates the layers of spatiotemporal perceptions. τ_{Total} indicates the total loss of our frameworks. $C-CHA$ indicates the cross-channel attention component, *GNPA* indicates the Group-Norm position attention component. $Conv_{7\times7}(\cdot)$ indicates that the convolutional operation of the kernel size is 7×7, $MP3\times3(\cdot)$ indicates that the max pooling operation of the kernel size is $3\times3.Conv_{1\times1}(\cdot)$ indicates that the convolutional operation of the kernel size is $1\times1.GN$ indicates the Group-Norm operation. × indicates the elementwise product operate

$x^{(1)} \in R^{C\times H\times W}$ and $x^{(2)} \in R^{C\times H\times W}$, respectively, where H, W, C indicates the height width and channel, , and the image size of the inputs is 256 × 256 × 3. The feature information of the output via the feature extraction module is $f^{Spatial}$,$f^{Temporal} \in R^{C\times H\times W}$, where C indicates the channel dimension. The spatiotemporal feature information is refined to attention feature maps $y^{(1)}$ and $y^{(2)}$ via a spatiotemporal perceptions module. However, the module is mainly composed of efficient channel attention guided squeeze-and-excitation. Then, we resize the optimization feature information to the size of the input remote sensing images. Meanwhile, we will calculate the distance of each pixel pair in the corresponding feature maps and archive a corresponding distance map ζ in the proposed optimization update.

Spatiotemporal feature extraction via DenseNet

In the past ten years, convolutional neural networks and improved convolutional neural networks [19] have been widely used in urban and rural perception tasks relying on their strong learnability, which is to expand the single dimension of traditional spatial structure to include morphological structure and intention type (City Intention Classification) and Intention Evaluation (Disaster Assessment) [22, 27] with other dimensions to extract better detailed information. Compared with traditional handmade or manual field survey methods, the method based on the convolutional neural networks not only has a higher efficiency, but also shows a stronger performance. To obtain better detailed information and different scales of spatiotemporal information, we introduce DenseNet to model urban and rural images in different phases, while using it as a feature extractor to capture multiscale spatiotemporal information to further enhance perception.

Due to the large differences in the socioeconomic environments of cities and villages and their different distribution states, such as landscapes, landmark buildings, public places, and cultural function areas, there is a strong spatial correlation between them. At the same time, there are interclass or intraclass differences in a certain spatial dimension, and the multiscale DenseNet can highlight these differences through features, such as feature multiplexing and information cross-layer connection, which can better represent high-level information. However, the original DenseNet was mainly used for image classification tasks and was directly used to capture the feature information of urban and rural socioeconomic environments (including buildings, roads, etc.). Therefore, we remove the final fully connected layer and use different scales of densely connected blocks [34] to obtain multiscale information on these intentional targets. DenseNet [35] high-level information is semantically accurate, but it cannot effectively determine the position of the intended target; the position of the intended target in the same area image cannot be determined in different time phases. The low-level information contains a wealth of physical structure and appearance details. To this end, we fuse the high-order and low-level layers in the spatial dimension to generate more refined representations. We also quantify and evaluate the intentional goals of cities and villages from different angles. It is worth noting that both the temporal branch and spatial branch use the multiscale DenseNet as the feature extractor. Assume that each densely connected convolutional block (Dense Block) is composed of l layers; however, the extraction process of multiscale spatiotemporal features $x^{(1)}_{S:l}$ and $x^{(2)}_{S:l}$ can be expressed as

$$\begin{cases} x^{(1)}_{S:l} = H_l([x^{(1)}_0, x^{(1)}_1, \ldots, x^{(1)}_{l-1}]),\ x^{(1)}_{MS:l} \in R^{C\times H\times W} \\ x^{(2)}_{S:l} = H_l([x^{(2)}_0, x^{(2)}_1, \ldots, x^{(2)}_{l-1}]),\ x^{(2)}_{MS:l} \in R^{C\times H\times W} \end{cases} \tag{1}$$

where, l indicates the number of layers and $l \geq 1$. $H_l(\bullet)$ indicates the operate of DenseNet-121. $x^{(1)}_0$=$T^{(1)}$, $x^{(2)}_0$ =$T^{(2)}$. S indicates multi scale information.

Spatiotemporal perceptions with cross-channel interaction attention

To further perceive the changes in the socioeconomic environments of cities and villages in recent years, to strengthen the dependence and location information between the same intentional target in different time phases and to improve the network's perception of the intentional target, we design a squeeze-and-excitation (SE) [36] enhanced channel attention the force module captures of the rich global spatiotemporal relationships among the intentional individuals throughout the entire time and space. It also establishes effective long short-term dependencies to highlight the perception of subtle changes and temporal and spatial characteristics, while providing subsequent urban and rural planning, disaster evaluation, and statistics. In addition, this model intends to provide reliable theoretical support for other tasks. The specific intention perception can be divided into the following steps:

> *Step 1* We first use the multibranch DenseNet-121 to obtain multiscale spatiotemporal information $f^{Spatial}$ and $f^{Temporal}$, which is a different Dense Block output of different scale information and is defined as $f^{Temporal}_s$ and $f^{Spatial}_s$, where $s \in S = 1, 2, 3, 4$. We can also think that $f^{Temporal}_1$ is equal to the output features of Dense block-1 (see Fig. 1). We first fused the captured multibranch spatiotemporal information and denoted it as.

$$\begin{cases} f_{Temporal} = f_S^{Temporal} = Conv_{1\times 1}(f_1^{Temporal}, \cdots, f_4^{Temporal}) \\ f_{Spatial} = f_S^{Spatial} = Conv_{1\times 1}(f_1^{Spatial}, \cdots, f_4^{Spatial}) \end{cases} \tag{2}$$

where $f_{Temporal}$ and $f_{Spatial}$ indicate the multibranch spatiotemporal information. $Conv_{1\times 1}$ indicates the operation of convolution, and the kernel size is 1×1. In addition, according to Equations 1 and 2, feature information of different scales can be expressed as

Step 2. We divide $f_{Temporal}$ into a γ subspace along the channel of temporal feature maps and $f_{Spatial}$ into a ε subspace along the channel of spatial feature maps. Finally, these subspaces are defined as.

$$\begin{cases} f_{Temporal} = [f_{Temporal_1}, f_{Temporal_2}, \cdots, f_{Temporal_\gamma}] \\ f_{Spatial} = [f_{Spatial_1}, f_{Spatial_2}, \cdots, f_{Spatial_\varepsilon}] \end{cases} \tag{3}$$

Then, the specific temporal semantics information of urban and village intended objects via each subspace $f_{Temporal_i} \in R^{\frac{C}{\gamma}\times H\times W}$ generate a corresponding coefficient. The structure of the spatial branch is similar to that of the temporal branch, namely, $f_{Spatial_i} \in R^{\frac{C}{\varepsilon}\times H\times W}$ and ε is the subspace.

Step 3. To make the module more portable and more conducive to the statistics of global information, we use the spatiotemporal information of the urban and rural intentional targets captured by the temporal branch as the input of the cross-channel attention (CHA) component, and under the condition of no dimensionality reduction, cross dimensionality embedding is performed on the intentional object. The cross-dimensionality embedding of urban and village intended objects via the cross-channel attention component is shown.

$$\begin{cases} f_{\gamma 1}^{CH} = \delta(\frac{W_1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\Psi) \cdot f_{Temporal_{\gamma 1}} \\ \Psi = f_{Temporal_{\gamma 1}}(i,j) + b_1 \end{cases} \tag{4}$$

However, the feature information captured by the spatial branch is used as the input of the Group-Norm position attention (GNPA) [37] component to determine the changing position of the urban and rural intentional target, which complements the output information of the cross-channel attention component (CHA). The output information of the GNPA component can be obtained with.

$$f_{\varepsilon 1}^{GNPA} = \delta(W_1' \cdot GN(x_{Spatial_{\varepsilon 1}}) + b_1') \cdot f_{Spatial_{\varepsilon 1}} \tag{5}$$

where $W_1 \in R^{\frac{C}{2\gamma}\times H\times W}$ and $W_1' \in R^{\frac{C}{2\varepsilon}\times H\times W}$ indicates the weighting factor of different components. $b_1 \in R^{\frac{C}{2\gamma}\times H\times W}$ and $b_1' \in R^{\frac{C}{2\varepsilon}\times H\times W}$ indicate the bias of different branch component. $\delta(\cdot)$ indicates the activities functional *ReLU* .

Meanwhile, to ensure efficiency, reliability, and help from effective cross-channel interaction between local and global information, the frequency band matrix W_γ is used to further improve the cross-channel attention (CHA) component, and it can be expressed as

$$W_\gamma = \begin{bmatrix} w^{1,1} & \cdots & w^{1,\gamma} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,\gamma+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-\gamma+1} & \cdots & w^{C,C} \end{bmatrix} \tag{6}$$

Where, w_γ indicates the weighting factor.

Step 4 We share these spatial and temporal branches to make the size of the feature maps the same as the initial inputs. The aggregation processing are denoted as .

$$f_{STPM} = \left[f_{\gamma 1}^{CH}, f_{\varepsilon 1}^{GNPA}\right], f_{STPM} \in R^{C\times H\times W} \tag{7}$$

We use different branches to capture the characteristic information of the urban and rural intentional targets to not only obtain better high-level information, but also obtain appearance details, establish a dependency relationship in the spatial and temporal dimensions, and further strengthen the relationship between humans and the urban and rural intentional targets. Interactivity improves the ability of follow-up applications.

Optimization

To further improve the representations and perceptibility of this spatiotemporal information for urban and village intention object changes, we present a loss functional of reconstruction. The loss functional are indicated as

$$\begin{cases} \tau_{Total} = \alpha\tau_{spatial} + \beta\tau_{temporal} \\ \alpha = 1 - \beta \end{cases} \tag{8}$$

Where α and β is a learnable balance factor.

For the spatial and temporal branches, we use the binary cross entropy loss (BCELoss) and cross entropy loss, namely, $\tau_{spatial}$ and $\tau_{temporal}$.

$$\begin{cases} \tau_{spatial} = -\frac{1}{N}\sum_{n=1}^{N}[y_n^{(1)}log(z_n) + (1 - y_n^{(1)})log(1 - z_n^{(1)})] \\ \tau_{temporal} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{u=1}^{U} y_{nu}^{(2)}log(z_{nu}^{(2)}) \end{cases} \tag{9}$$

where N indicates the total number of samples, $y_n^{(1)}$ is the category of the n-th sample, $z_n^{(1)}$ is the predicted value of the n^{th} samples, $u \in U$ indicates the number of categories, and $z_{nu}^{(2)}$ indicates the probability that the n^{th} sample belongs to category u.

In summary, we can better perceive changes in the content of the urban and rural intentional targets in this way, achieve as much automated processing of the content as possible, and improve the ability and efficiency of emergency response after disasters. The proposed multibranch networks for exploring spatiotemporal changes in cities and villages are shown in Algorithm 1.

```
Algorithm 1: Spatiotemporal changes in cities and villages by multibranch
networks
Input:Assume that the remote sensing images before and the space before and after the
$x^{(1)} \in R^{C\times H\times W}$ and $x^{(2)} \in R^{C\times H\times W}$, where $C, H, W$ indicates the channel, height and
width, respectively.$l \in L$ indicates the layers of multibranch networks;
for $s$=1 to $S$ do
    Calculate the multiscale spatial-temporal information $f_{Temporal}$ and $f_{Spatial}$ by the
      backbone component of the DenseNet. The detailed visible Eq. (1) and Eq.(2). ;
    The spatiotemporal preceptions $f^{CH}$ and $f^{GNPA}$ are achieved via cross-channel
      interaction attention networks. Such as Eq. (3)   Eq. (6);
    The aggregation information $f_{STPM}$ by the STPM component is given by Eq. (7);
end
output:optimize the training by reconstructing $\tau_{Total}$ and change the spatiotemporal results of
  cities and villages, according to Eq. (8) and Eq. (9);
```

Experimental discussion and analysis of the spatiotemporal perceptions

In the sections, we describe our perception results of urban and village intention objects in detail and provide a discussion and analysis.

Data preparation and processing

Because there is no database specifically used to perceive the changes in the urban and rural intentional targets, we screen public baseline datasets, such as LEVIR-CD and SZAB.

LEVIR-CD: [38] The dataset has a total of 637 1024*1024 remote sensing images and mainly describes the changes in urban and rural buildings in 20 different areas of several cities in Texas, USA, between 2002 and 2018, mainly concentrating on the growth of the various types of buildings (such as villas, high-rise apartments, small garages and large warehouses) in cities and villages.

SZAB: [39] The datasets are called the SZTAKI-Air-Change-Benchmark and contain 13 pairs of aerial images with a size of 952x640 pixels and a spatial resolution of 1.5 m. It mainly includes new urban areas, building construction, planting a large number of trees and new cultivated land.

However, to better perceive the changes in the urban and rural intentional environments and provide more reliable experimental support for subsequent urban planning, intention type or disaster evaluation, we preprocessed this initial data to ensure that the processed dataset was suitable for urban and rural areas. The description of a socioeconomic environment is more comprehensive, and it is also more suitable for urban and rural perception tasks.

Training configuration

To achieve a better perception effect of city and village intention objects by training our proposed frameworks, we conduct a sequence of initial settings for the frameworks and enhance these datasets by augmentation methods. Meanwhile, augmentation can also be effective to compensate for the lack of urban and rural content data, such as rotation, noise, color change, etc. The processing of the datasets are shown as Fig. 2.

For the network structure of our present spatiotemporal perception frameworks, the scale is set as $s \in \{S = 1, 2, 3, 4\}$, the growth rate for the DenseNet-121 is $k = 32$, and the learning rate is set as $1e - 4$. The *Dropout* is 0.5 and the epoch is set as 600. However, to further ensure the effectiveness of training for our frameworks, we force the input remote sensing image size to be cropped to 256 × 256. The datasets are divided into three subsets: training (40%), testing (60%) and validation (10%).

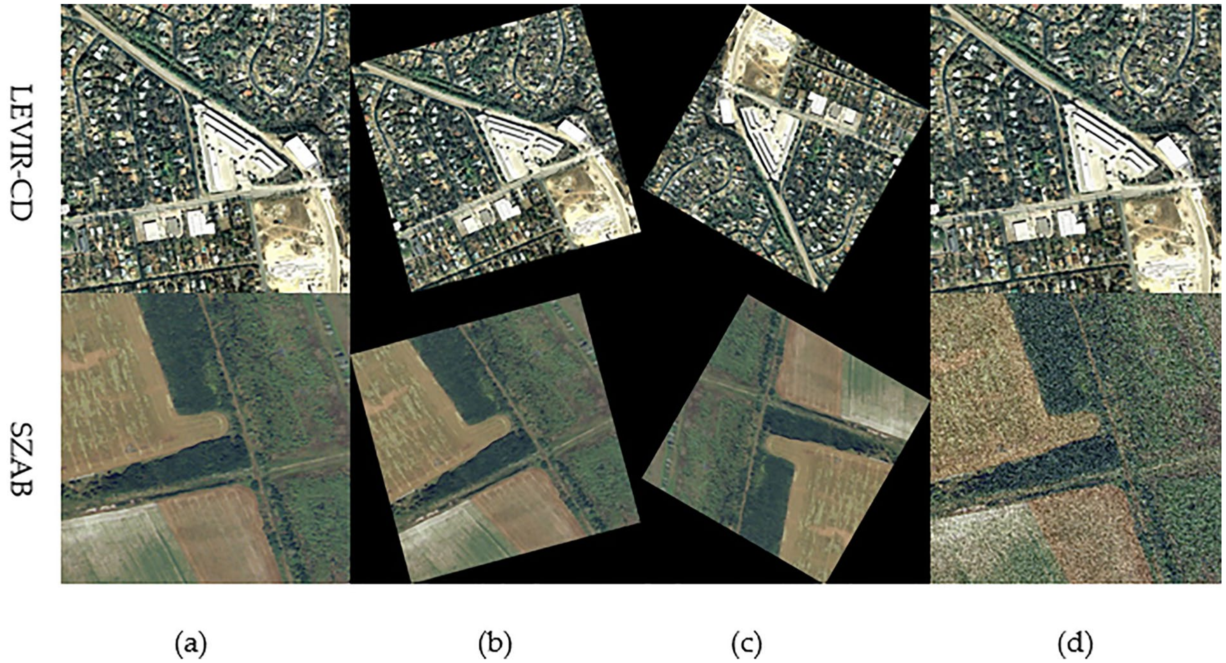

Fig. 2 The augmentation results of the LEVIR-CD and SZAB baseline datasets. **a** indicates the original image of urban areas and villages. **b** and **c** Indicate that they are rotated by 15 degrees and 150 degrees, respectively. **d** Indicates that Gaussian noise is added

Evaluation coefficient

To ensure the consistency and validity of the experimental results, we use multiple evaluation coefficients, such as precision (P), recall (R), F-score (F1), average area (AR), parameter quantification (PQ) and time, where "time" indicates the run time of each batch size to test and verify our experiments. The calculation process of the evaluation index is shown in the following equation.

$$\begin{cases} P(Precision) = \frac{TP}{TP \times FP} \\ R(Recall) = \frac{TP}{TP \times FN} \\ F(F_1) = \frac{2 \times P \times R}{P+R} \\ AR(AverageAreas) = \frac{PCR}{TR} \end{cases} \tag{10}$$

where TR indicates the total area of remote sensing images and PCR indicates the change area via predic.

Experimental results of the different methods

To demonstrate the effectiveness of our proposed spatiotemporal perception framework, it also helps to collect information on the environments of urban and rural cities, and improve the responsiveness of tasks, such as urban planning, disaster evaluation and intention type judgment. Compared with other advanced perception frameworks, we tested and verified two datasets, LEVIR-CD and SZAB, with precision, recall and F-score as evaluation indicators. Meanwhile, we will give the change area of the intention content in the urban and rural images, namely, AR. The experimental results of the different methods are shown in Table 1.

According to Table 1, we can draw the following conclusions:

1. The perception framework we propose achieves the best results in a variety of evaluation indicators. The main reason may be that the multiscale spatiotemporal information extracted by the dual-branch DenseNet-121 is used to describe the urban and rural targets in detail from different angles and different levels. At the same time, attention is used to aggregate multilevel information, which further strengthens the use of the detailed information and strengthens the interaction between the temporal and spatial information.

   In addition, the perception framework we propose is also very competitive in terms of perception. The parameter amount and time efficiency are 18.14 M and 11.95 s, respectively, which is 4.14 s higher than the KPCAMNet method in efficiency. The possible reason for this is that in our proposed perception framework, the squeeze excitation component uses a

Table 1 The perception results of our proposed frameworks: where *AR* indicates the percentage before and after the average area change. *PQ* indicates the parameter quantification of the methods. *Time* indicates the run time of each batch size." –" means equal values for the same model

Datasets	Model	P(%)	R(%)	F(%)	AP(%)	PQ(M)	Time(s)
LEVIR-CD	VGG-LR	63.54	65.19	64.35	12.33	4.12	2.01
	ChangeNet	64.98	67.56	66.24	13.08	6.19	3.81
	FDCNN	67.49	68.95	68.21	14.59	6.01	4.92
	CD-UNet++	68.55	70.07	69.30	15.38	5.54	4.53
	UNetLSTM	69.28	71.05	70.15	17.18	12.01	7.32
	SRCDNet	74.83	80.62	77.62	20.59	10.11	6.07
	ESCNet	77.07	84.16	80.45	23.44	17.22	11.88
	FGCN [29]	78.51	79.44	78.97	24.94	19.72	14.43
	PToP CNN [26]	79.92	81.57	80.74	26.26	20.02	15.84
	KPCAMNet	81.56	86.07	83.75	27.48	21.49	16.09
	Ours	84.38	92.04	88.04	31.43	18.14	11.95
SZAB	VGG-LR	33.24	35.52	34.34	5.19	–	–
	ChangeNet	34.28	37.16	35.66	5.86	–	–
	FDCNN	35.12	38.05	36.52	6.27	–	–
	CD-UNet++	36.94	38.99	37.93	6.88	–	–
	UNetLSTM	37.52	39.65	38.56	7.22	–	–
	SRCDNet	40.69	42.28	41.47	9.44	–	–
	ESCNet	42.14	44.37	43.23	10.87	–	–
	FGCN [29]	43.49	45.74	44.57	11.04	–	–
	PToP CNN [26]	44.23	46.07	45.13	11.26	–	–
	KPCAMNet	44.52	46.01	45.25	11.32	–	–
	Ours	46.15	64.27	53.72	13.49	–	–

reduced number of parameters without reducing the perception accuracy.

2. Compared with the convolutional neural network methods (VGG-LR, ChangeNet and FDCNN), the U-Net method (CD-UNet++ and UNetLSTM) achieves better perceptual effects. For example, on the LEVIR-CD dataset, the perceptual performance of the UNetLSTM is improved by 1.79% (P), 2.1% (R), and 1.94% (F) compared to the FDCNN. The possible reason is that when the UNet encodes and decodes urban and rural targets, it better captures the detailed semantics of the target, and the description of the target is more detailed.
   Compared with the CD-UNet++, the UNetLSTM achieves a better perceptual performance. The main reason is that it not only uses the U-Net to encode and decode the local features of urban and rural targets but also uses the LSTM to describe the global semantics of the target connotation. Expressing urban and rural goals from two perspectives, local and overall, forms a complementarity. Compared with other CNN-based perception methods, the VGG-LR achieves the lowest effect. The main reason for this is that the framework only uses the VGG-16 to extract local features of urban and rural targets, and loses a large amount of detailed information.
3. Compared with other methods, such as the ESCNet, SRCDNet and UNetLSTM. The two perception frameworks, the FGCN and the PtoP CNN, have achieved better performance. The main reason is that different branches are used to model the local and global semantics of the target, which forms an interaction and complementarity between the global and local semantics, improving the feature information pair and the ability to perceive subtle changes.
4. Compared with the perception methods based on the CNN and the U-Net, the SRCDNet, ESCNet and KPCAMNet have strong competitiveness. For example, on the SZAB data, the KPCAMNet method has increased by 7.0%, 6.36% and 6.69%, respectively, compared to the UNetLSTM. The main reason for this is that the KPCAMNet uses multiscale information and simultaneously uses attention to refine the multiscale information, filtering out redundant information. In addition, the number of participants in the training of the perceptual framework we propose is also small.

Table 2 The perception results of our proposed frameworks shows where *AR* indicates the percentage before and after the average area change

Datasets	Model	P(%)	R(%)	F(%)	AP(%)	PQ(M)	Time(s)
LEVIR-CD	Ours (non-STPM)	80.15	88.94	80.51	27.13	16.97	10.41
	Ours	84.38	92.04	88.04	31.43	18.14	11.95
SZAB	Ours (non-STPM)	43.22	63.31	51.37	12.46	–	–
	Ours	46.15	64.27	53.72	13.49	–	–

Experimental results of different components

To verify the impact of the different components on the overall performance of the proposed perception framework based on baseline data, such as the LEVIR-CD and the SZAB, different components were tested and demonstrated, and the experimental results and related analysis are given. The specific experimental results are shown in Table 2.

According to Table 2, we can draw the following conclusions:

Our proposed spatiotemporal perception framework achieves the best performance on the two public baseline datasets; the F-scores were 88.04% and 53.72%, respectively. The main reason for this is that the perception framework we design uses multibranch deep neural networks to first capture the deep semantics and shallow physical appearance information of the urban and rural intentional targets, while describing the intentional targets from different levels and scales. Second, to further establish a spatiotemporal dependence, interaction modeling between long- and short-term distances can be used to more accurately mark the position of the intentional target. At the same time, it highlights the difference between the intentional target class or the class and further improves the network's perception of the socioeconomic environment of the urban and rural areas. ability. In addition, we can also find that only using the DenseNet (Our(No-STPM)) for spatiotemporal information extraction can also achieve better performance, but compared to using the STPM module (Our), its F-score value is reduced by 7.47% and 2.35%, respectively.

At different times, the urban and rural intentions in the same area showed great changes, and the average areas of change were 31.43% and 13.49%, respectively. This shows that with the continuous development of the social economy, the urban and rural forms will also undergo massive changes. If artificial participation is used, it is time-consuming and labor-intensive to measure the changing area, and the method we provide can effectively improve the measurement efficiency; at the same time, it is more

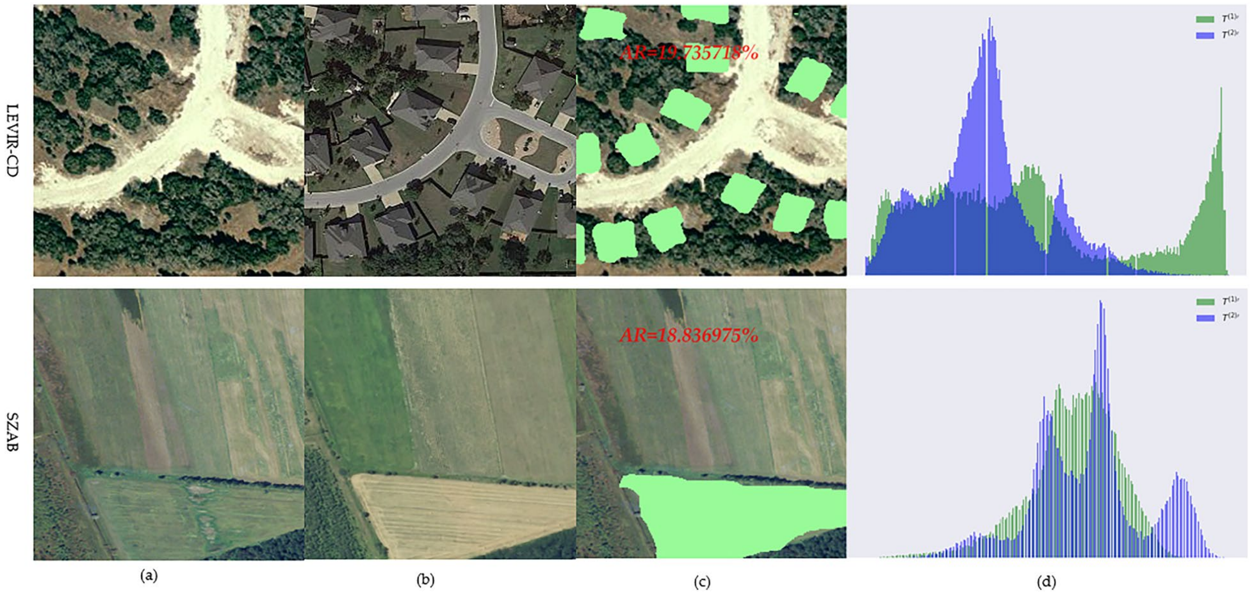

Fig. 3 The perceptions results of our present frameworks. **a** and **b** Indicate the image urban and rural areas before and after the change. **c** Indicates the perception results, where white represents the part of the perceived change. **d** Indicates the histogram

Table 3 Experiment results of different component

Model	P(%)	R(%)	F(%)	AP(%)	PQ(M)	Time(s)
Ours(No-CHA)	81.54	89.35	85.27	28.09	14.54	9.28
Ours(No-GNPA)	82.97	90.69	86.66	28.88	15.22	10.07
Ours	84.38	92.04	88.04	31.43	18.14	11.95

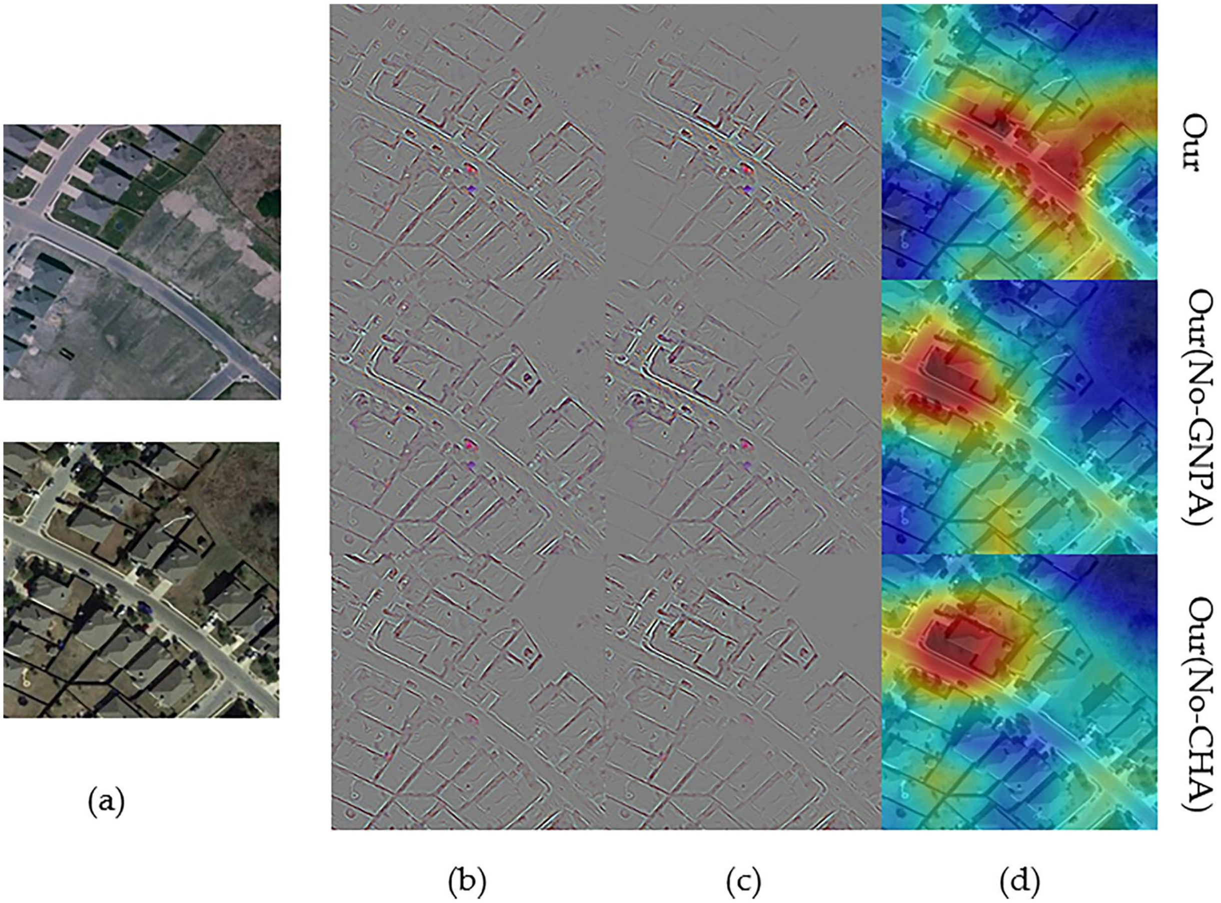

Fig. 4 The hotmaps of the different components. **a** Indicates the initial image of urban and rural areas before and after the change. **b** and **c** Indicate the spatiotemporal feature maps of the middle layers. **d** Indicates different components of the hotmap

accurate, providing a certain experimental basis for subsequent urban planning and assumptions.

To show the performance of our proposed spatiotemporal perception framework more intuitively, we give the perception effects of different regions, where the results are shown in Fig. 3.

Ablation studies

To further verify the influence of the different components on the proposed framework, experimental tests are carried out based on the LEVIR-CD datasets, and the relevant perception results and analysis are given. The perceptions results are shown in Table 3.

According to Table 3, we can find that the perception accuracy of using CHA (our(No-GNPA)) is obviously better than using GNPA (Ours(No-CHA)) and that its F value and AP are increased by 1.39% and 0.79%, respectively. This indicates that CHA's contribution to the network is higher than that of GNPA. The main reason for this may be that CHA captures more effective specific information and is more sensitive to urban and rural objects. However, to better show the impact of CHA and

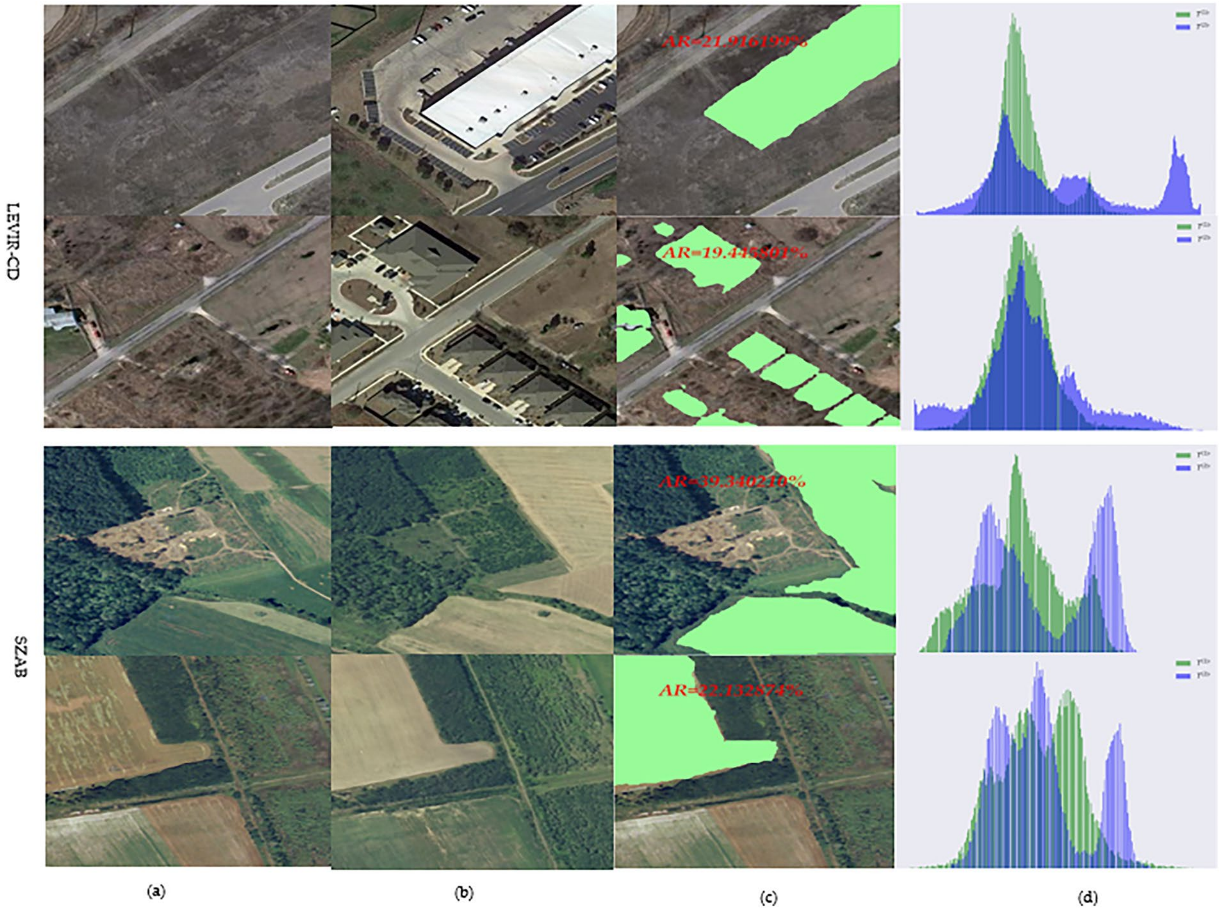

Fig. 5 The perception results of dour proposed frameworks. **a** and **b** Indicate the image urban and rural areas before and after the change. **c** Indicates the perception results, where the white part of **c** represents the perceived change part. **d** Indicates the histogram of urban and village intention objects

GNPA components on the overall frame performance, we have provide a visual hotmap of different components. The hotmap are denoted in Fig. 4.

According to Fig. 4, we can obviously see that the two components are used in conjunction to form information complementarity, which can better express the urban and rural intentional targets and, at the same time, can perceive subtle changes. Because CHA uses cross-channel interaction to capture the specific semantics of the urban and rural intentional targets, GNPA can better locate the target's location, and their collaborative work can establish a more effective dependence.

The discussion of the results

To show that the proposed perception framework can effectively detect the socioeconomic environments of urban and rural locations, forms, and infrastructure, while contributing to various tasks, such as disaster evaluation and intention type statistics, we show the perception results of multiple intentional targets. The result is shown in Fig. 5.

Conclusions and next studies

In this paper, we perceive the changes in the socioeconomic environments of urban and rural areas, and present an exploration of spatiotemporal changes in cities and villages through remote sensing using multibranch networks. The perception framework not only effectively captures the multiscale spatiotemporal information of the intended target, but also uses STPM to capture the long-term spatiotemporal correlation. The intended target is described from multiple perspectives, such as high-level semantics and low-level appearance to learn more effective embeddings. In addition, the interaction between time and space information is strengthened, and this characteristic information is gradually refined during the training process, which is helpful for urban planning and construction and disaster response. The final perception results show that our proposed perception framework has a good performance.

Although the framework has achieved a good perceptual performance, the perceptual effect of the intentional targets with large scale changes (the same target at different moments or on different remote sensing images, the physical appearance of the intended target, such as the shape and size of the intended target changes greatly) is poor and needs to be improved. Therefore, in future work, we will introduce concepts like as superscale blocks to develop a simpler and more effective semantic framework. At the same time, we will further improve the attention network to guide the perception framework to explore large-scale changes. Finally, we learn the important characteristics of the urban and rural socioeconomic areas.

Abbreviations

GNPA: Group-Norm position attention; CHA: Cross-channel attention component; STPM: Spatio temporal perceptions.

Acknowledgements

Not applicable.

Authors' contributions

MZ:Conceptualization, Methodology,Software:Programming, implementation of the computer code and supporting algorithms ,Writing- Original draft preparation. YT:Writing-Review and Editing,Supervision,Funding acquisition. Both authors read and approved the final manuscript.

Funding

This work is supported by National Natural Science Foundation of China–Youth Science Fund (Grant Number:51308427).

Availibility of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan 430000, China. [2]Wuhan Natural Resources and Planning Bureau, Wuhan 430000, China.

Received: 28 July 2021 Accepted: 13 September 2021
Published online: 26 September 2021

References

1. Naik N, Philipoom J, Raskar R, Hidalgo C. Streetscore-predicting the perceived safety of one million streetscapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014. pp. 779–785.
2. Gebru T, Krause J, Wang Y, Chen D, Deng J, Aiden EL, Fei-Fei L. Using deep learning and google street view to estimate the demographic makeup of the us. arXiv preprint arXiv:1702.06683; 2017.
3. Li X, Cai BY, Ratti C. Using street-level images and deep learning for urban landscape studies. Landscape Arch Front. 2018;6(2):20–30.
4. Zhou H, Liu L, Lan M, Zhu W, Song G, Jing F, Zhong Y, Su Z, Gu X. Using google street view imagery to capture micro built environment characteristics in drug places, compared with street robbery. Comput Environ Urban Syst. 2021;88:101631.
5. Wegner JD, Branson S, Hall D, Schindler K, Perona P. Cataloging public objects using aerial and street-level images-urban trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016; pp. 6014–6023.
6. Obeso AM, Benois-Pineau J, Vázquez MG, Acosta AR. Saliency-based selection of visual content for deep convolutional neural networks. Multimedia Tools Appl. 2019;78(8):9553–76.
7. Morbidoni C, Pierdicca R, Paolanti M, Quattrini R, Mammoli R. Learning from synthetic point cloud data for historical buildings semantic segmentation. J Comput Cult Herit. 2020;13(4):1–16.
8. Hu Y, Manikonda L, Kambhampati S. What we instagram: A first analysis of instagram photo content and user types. In: Eighth International AAAI Conference on Weblogs and Social Media; 2014.
9. Hochman N, Manovich L. Zooming into an instagram city: Reading the local through social media. First Monday; 2013.
10. Jett J, Senseney M, Palmer CL. Enhancing cultural heritage collections by supporting and analyzing participation in flickr. Proc Am Soc Inform Sci Technol. 2012;49(1):1–4.
11. Li J, Xie X, Zhao B, Xiao X, Qiao J, Ren W. Identification of urban functional area by using multisource geographic data: A case study of zhengzhou, china. Complexity **2021** (2021).
12. Bose A, Chowdhury IR. Monitoring and modeling of spatio-temporal urban expansion and land-use/land-cover change using markov chain model: a case study in siliguri metropolitan area, west bengal, india. Model Earth Syst Environ. 2020;6(4):2235–49.
13. Liu X, Tian Y, Zhang X, Wan Z. Identification of urban functional regions in chengdu based on taxi trajectory time series data. ISPRS Int J Geo-Inform. 2020;9(3):158.
14. Tardioli G, Kerrigan R, Oates M, O'Donnell J, Finn DP. Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach. Build Environ. 2018;140:90–106.
15. Gadal S, Ouerghemmi W. Identification of urban objects using spectral library combined with airborne hyperspectral imaging. In: 4ème Colloque du Groupe Hyperspectral de la Société Française de Photogrammétrie et Télédétection (SFPT-GH); 2016.
16. Manzoni M, Monti-Guarnieri A, Molinari ME. Joint exploitation of spaceborne sar images and gis techniques for urban coherent change detection. Remote Sens Environ. 2021;253:112152.
17. Llamas J, M Lerones P, Medina R, Zalama E, Gómez-García-Bermejo J. Classification of architectural heritage images using deep learning techniques. Appl Sci 2017; 7(10), 992.
18. Chen Q, Cheng Q, Wang J, Du M, Zhou L, Liu Y. Identification and evaluation of urban construction waste with vhr remote sensing using multi-feature analysis and a hierarchical segmentation method. Remote Sens. 2021;13(1):158.
19. Attari N, Ofli F, Awad M, Lucas J, Chawla S. Nazr-cnn: Fine-grained classification of uav imagery for damage assessment. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 50–59 (2017). IEEE
20. Vetrivel A, Gerke M, Kerle N, Nex F, Vosselman G. Disaster damage detection through synergistic use of deep learning and 3d point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. ISPRS J Photogramm Remote Sens. 2018;140:45–59.
21. Hamdi ZM, Brandmeier M, Straub C. Forest damage assessment using deep learning on high resolution remote sensing data. Remote Sens. 2019;11(17):1976.
22. Li X, Caragea D, Caragea C, Imran M, Ofli F. Identifying disaster damage images using a domain adaptation approach. In: Proceedings of the 16th International Conference on Information Systems for Crisis Response And Management (2019).
23. Meng L, Wen K-H, Zeng Z, Brewin R, Fan X, Wu Q. The impact of street space perception factors on elderly health in high-density cities in macau-analysis based on street view images and deep learning technology. Sustainability. 2020;12(5):1799.

24. Kim D, Kang Y, Park Y, Kim N, Lee J. Understanding tourists' urban images with geotagged photos using convolutional neural networks. Spatial Inform Res. 2020;28(2):241–55.
25. Chen M, Arribas-Bel D, Singleton A. Quantifying the characteristics of the local urban environment through geotagged flickr photographs and image recognition. ISPRS Int J Geo-Inform. 2020;9(4):264.
26. Jayasuriya M, Arukgoda J, Ranasinghe R, Dissanayake G. Localising pmds through cnn based perception of urban streets. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 6454–6460 (2020). IEEE
27. Wang T, Tao Y, Chen S-C, Shyu M-L. Multi-task multimodal learning for disaster situation assessment. In: 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 209–212 (2020). IEEE
28. Agarwal M, Leekha M, Sawhney R, Shah RR. Crisis-dias: Towards multimodal damage analysis-deployment, challenges and assessment. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 346–353 (2020).
29. Zhao X, Tao R, Li W, Philips W, Liao W. Fractional gabor convolutional network for multisource remote sensing data classification. IEEE Trans Geosci Rem Sens. 2021.
30. Zhao X, Tao R, Li W, Li H-C, Du Q, Liao W, Philips W. Joint classification of hyperspectral and lidar data using hierarchical random walk and deep cnn architecture. IEEE Trans Geosci Rem Sens. 2020;58(10):7355–70.
31. Zhang M, Li W, Tao R, Li H, Du Q. Information fusion for classification of hyperspectral and lidar data using ip-cnn. IEEE Trans Geosci Rem Sens. 2021.
32. Zhang M, Li W, Du Q, Gao L, Zhang B. Feature extraction for classification of hyperspectral and lidar data using patch-to-patch cnn. IEEE Trans Cybern. 2018;50(1):100–11.
33. You H, Tian S, Yu L, Lv Y. Pixel-level remote sensing image recognition based on bidirectional word vectors. IEEE Trans Geosci Rem Sens. 2019;58(2):1281–93.
34. Cheng Q, Xu Y, Fu P, Li J, Wang W, Ren Y. Scene classification of remotely sensed images via densely connected convolutional neural networks and an ensemble classifier. Photogrammetr Eng Remote Sens. 2021;87(4):295–308.
35. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017; pp. 4700–4708.
36. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018; pp. 7132–7141.
37. Zhang Q-L, Yang Y-B. Sa-net: Shuffle attention for deep convolutional neural networks. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2235–2239 (2021). IEEE
38. Chen H, Shi Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sens. 2020;12(10):1662.
39. Benedek C, Szirányi T. Change detection in optical aerial images by a multilayer conditional mixed markov model. IEEE Trans Geosci Remote Sens. 2009;47(10):3416–30.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.