Exploring Spatiotemporal Change of City and Village from Remote Sensing Using Multi-Branch Networks

Mengqi Zhao (zhaomengqi@whut.edu.cn)
Wuhan University of Technology https://orcid.org/0000-0002-2029-7759
Yan Tian
Wuhan University of Technology School of Civil Engineering and Architecture

Research article

Keywords: Intentional target, spatiotemporal changes, multi-scale spatiotemporal information, cross-channel attention, information interaction

Posted Date: August 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-727020/v1

Undefined 1 (2009) 1–5
IOS Press

Exploring Spatiotemporal Change of City and Village from Remote Sensing Using Multi-Branch Networks

Zhao Mengqi [a,*,**],
Tan Yan [b]

[a] *School of Civil Engineering and Architecture, Wuhan University of Technology, WuHan, China*
E-mail: zhaomengqi@whut.edu.cn
[b] *School of Civil Engineering and Architecture, Wuhan University of Technology, WuHan*
Natural Resources and Planning Bureau, WuHan China

Abstract. With the rapid development of social economy, the urban and rural environment, form and infrastructure have also undergone earth-shaking changes. As a gathering place for human activities, urban and rural areas play a vital role in the interaction between humans and society. If traditional machine learning methods are used to perceive changes in the intentional connotation of urban and rural areas, it is easy to ignore the detailed information of the intentional target. At the same time, the perception accuracy needs to be improved. Therefore, the deep neural network in this paper proposes a way to perceive the temporal and spatial changes of urban and rural intentional connotations from the perspective of remote sensing. The framework first uses the multi-branch DenseNet to capture the multi-scale spatiotemporal information of the intended target, and embeds it with high-level semantics and low-level physical appearance information. Secondly, a multi-branch cross-channel attention module is designed to perform refined and integrated processing of multi-level spatiotemporal information to realize the aggregation of multi-branch information. Finally, it was tested and verified on two different time data sets, and good performance was achieved.

Keywords: Intentional target, spatiotemporal changes, multi-scale spatiotemporal information, cross-channel attention, information interaction

1. Introduction

With the continuous development of social economy, human living standards have also undergone tremendous changes, and cities and villages, as gathering places for human social interaction and activities,have also undergone tremendous changes in recent years. In addition, cities and villages not only reflect people's lifestyles as important representatives of human activities, but their changes also affect people's physical and 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 economy in depth. At the same time, it can also effectively judge whether it is necessary to further improve the infrastructure construction, and it is of great significance for improving the quality of urban and rural spaces and shaping the characteristic style.

In recent years, with the application of computer intelligent interpretation technology in many fields such as natural language processing (NLP), image classification, object detection (OD), etc., it has provided a new way of thinking for the exploration of City and Village environmental changes (including buildings, infrastructure and heritage, etc.).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,

*Corresponding author. E-mail: zhaomengqi@whut.edu.cn
**Do not use capitals for the author's surname.

and environment of a certain area, such as using digital models to build urban forms and urban environments. Model to obtain useful information. However, with the complexity of urban and rural environments, its digital modeling is also difficult to meet the increasing application requirements. At the same time, affected by the exponential growth of data under the condition of big data, it is difficult to obtain effective detailed information with this kind of simulation modeling method. Meanwhile, the manner of digital modeling usually over-simplify not available for all studies including the changes of infrastructure. Meanwhile, it neglect the city and villages netural landscapes and the effective is poor.

However, in order to obtain more detailed information from relevant data to effectively represent and simulate urban and rural environmental changes, many machine learning and deep learning algorithms have been developed. For instance, Naik, Nikhil and Philipoom et.al [1] to rate the safety, resident wealth and vitality index of the block, the data collected from Google will be input into the machine learning model for modeling, and new neighborhoods semantic information will be generated. Gebru, Timnit and Krause et.al [2] proposed a deep learning methods 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. Whereafter, Xiaojiang LI and Bill Yang CAI [**?**] present a urban landscape studies methods by combinations deep convolution neural networks (DCNN) and street-level images, and recognise the different urban features from these street-level images accurately. Meanwhile, Considering that machine learning and deep learning methods also have strong modeling ability for complex and large scale data, and apply 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 [**?**][5] [6]. Obeso, A Montoya and Benois-Pineau et.al [7] to addresses the classification problem of Mexican cultural heritage adopt a deep convolution netural networks methods to training and predict visual attention in natural images.Morbidoni, Christian and Pierdicca et.al [8] proposed a novelty methods of learning from synthetic point cloud data for historical buildings semantic segmentation, where the mainlyto provide a first assessment of the use of synthetic data to drive convolution netural networks based semantic segmentation in the context of historical buildings.

Although the above methods detect urban and rural form and environment to a certain extent, they basically focus on segmentation tasks such as content classification and recognition or buildings, and ignore the changes in the urban and rural spatiotemporal environment. At the same time, these methods ignore the image feature extraction process. The subtle changes in the target object and the poor perception of temporal and spatial semantics. Thus, we addresses these issues, we also explore changes in the urban and rural environment, form and infrastructure from the perspective of time and space. we present a novelty spatiotemporal perception methods, namely, exploring change of city and village from remote sensing with multi-branch networks. We aim to build visual spatiotemporal perception models that can be used to estimate environment, form and infrastructure changes of urban and villages, Meanwhile, also can be conducive to vigorously promote the development of social research and improving the lifestyles way of human.

Summary, the mainly work in the paper are following as:

- **Frameworks**: A methodology for exploring the spatiotemporal changes from remote sensing of city and villages environment, form and infrastructure aspect. The mainly aim to build relationship between haman visual and perceptions which can be done to understand the changes of social development, and improving the effective of statistic from society.
- **Technology**: we present a novelty perceptions frameworks using multi-branch networks.This method mainly uses a multi-branch attention network to model remote sensing images in the same area at different time periods, forming information sharing in time and space; secondly, through these characteristic information to perceive subtle changes in different targets in cities and villages, including targets Position, physical structure and geometric shape, etc., and further establish 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 data sets such as a and b, using preprocessing such as 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 organizational structure of the paper. In Section 2 we elaboration the related work of urban and villages environment, form and infrastructure et.al image perception. In Section 3 detailed describes our proposed perception frameworks. In Section 4 discusses and analyzes the processing of datasets and application. Then, we present a detailed results of experiments and describes the changes. In Section 5 conducive a briefly summary and possibilities of future work.

2. Related works

In the section, we would elaboration the related work of city and villages form, infrastructure et.al image perception. And the primary individe into traditional and deep neural networks of urban and villages environment form or infrastructure 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.

2.1. The traditional image perception of city and villages

Now, the image datasets have been widely using in many files of urban and villages research and planning in progress, such as, the mainly application files contains regional city systems, city and villages spatial structure, infrastructure service systems, transportation and travel and ollective activities of the people et.al.However, with the continuous development of society and 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 many algorithms to perceive urban and rural areas from different perspectives and archive effective information such as the form and environment of the plant. For instance,Hu, Yuheng and Manikonda, Lydia et.al [9] proposed a effective methods of typology analysis, namely, They use computer technology to check the content of different images and adopt a methods of clustering to judge the activity levels of different types users on Instagram. Hochman, Nadav and Manovich, Lev et.al [10] based on Instagram algorithms, a spatiotemporal pattern analysis method is designed to visualize the characteristics of image content from 13 different cities around the world, and make corresponding comparisons to further describe people's activities, culture, etc. However, in order to conducive interaction of users and existing images datasets and further Extended scale of these images datasets via social media, J Jett and M Senseney et.al [11] present a feedback framework for transferring user-generated information to institutional data providers, which can improving teh service scope of the datasets center, but the methods mainly using cultural heritage institutions that also can enhance collections by sharing content through popular web services. The above-mentioned methods mainly use some simple visual methods to analyze the images of cultural heritage, residents' living conditions and environment circulating on social media during the disaster, although realize quick and simple statistics to further expand the relevant database, but 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 the buildings. Such as, Liu, ZJ and Wang, J et.al [12] proposed a building extraction methods from high resolution remote sensing imagery, the methods mainly using a multi-scale object classification and probabilistic Hough transform which including color feature and texture feature information. Li, Jingzhong and Xie, Xiao et.al [13] to addresses the sustainable development problem of city and effective identification of urban functional areas, combinations with multisource geographic data which a quantitative measurement method for urban functional areas has been established. Berg, Alexander C and Grabler, Floraine et.al [14] proposed a methods of parsing images of architectural scenes, where the parsing contents including roof, vegetation, Windows, building boundaries and doors etc part by training feature information of color and texture. Bose, Arghadeep and Chowdhury et.al [15] take the Siliguri metropolitan area in West Bengal, India as the research object, proposes a novelty study methods of Markov Chain model and analyzing the spatial distribution of urban land. Liu, Xudong and Tian, Yongzhong et.al [16] to scientifically plan the urbanization layout and improve the utilization rate of land space, the urban functional areas are identified and analyzed from the perspective of data mining, and taxi trajectory data is 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 Mathias, Markus and Martinovic, Andelo et.al [17] present a automatic recognise methods of architectural style, and using the features extracted by the pyramid of the Gabor filter to train the support vector machine (SVM), but the methods ignore local temporal information in processing of features extracted. However, to addresses these limited, Chu, Wei-Ta and Tsai, Ming-Hung et.al [18] considering the spatial relationship between local and global features, proposed a high-level feature representation approach. Tardioli, Giovanni and Kerrigan, Ruth et.al [19] to evaluate the building energy in the city, a new method is proposed to identify building clusters and a data set of representative buildings is provided. At the same time, the method is mainly divided into three parts: building classification, building clustering and prediction. Gadal, Sébastien and Ouerghemmi, Walid et.al [20] considering that hyperspectral remote sensing images can describe surface objects and landscapes more accurately, a classification method based on urban target spectral database is proposed to detect and classify specific urban targets. Manzoni, Marco and Monti-Guarnieri, Andrea et.al [21] combining synthetic aperture radar (SAR) images and geospatial information systems, a simple and fast method to identify structural changes in buildings in urban environments is proposed, which can effectively evaluate small changes after disasters.

2.2. The deep learning image perception of city 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 details change 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,the technical of deep learning are widely used in tasks such as urban planning, urban building classification, and urban form perception. such as Llamas, Jose and M Lerones, Pedro et.al [22] present a novelty methods of the classification of images of architectural heritage with deep convolutional neural networks. The main objective of this article is 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 more effective search of the urban architectural heritage, and beneficial in the tasks of studying and interpreting the heritage asset in question. With rapid development of urban and villages, due to its wide distribution, construction waste is easily confused with the surrounding environment and difficult to be manually classified. At the same time, considering that traditional single-spectral feature analysis is difficult to extract and identify urban construction waste related information. Thus,Chen, Qiang and Cheng et.al [23] combined with the multi-feature analysis method of remote sensing images, a method for extracting urban construction waste information from the optimal VHR image combined with morphological index and hierarchical segmentation are proposed. Attari, Nazia and Ofli, Ferda et.al [24] to assess the extent of damage to urban and villages building structures after the disaster, combined with UAV imagery proposed a fine-grained classification method called Nazr convolution neural networks (Nazr-CNN) and conduct damage assessment. Vetrivel, Anand and Gerke, Markus et.al [25] 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 multi-core learning framework is used to combine the two kinds of information to achieve classification, and finally to perform damage detection on the building roof and other object. Subsequently, Hamdi, Zayd Mahmoud and Brandmeier et.al [26] present a forest damage assessment method with deep learning techniques, and the backbone network of the method is mainly U-Net.Although these methods have achieved good results in post-disaster 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. such as, Tien Nguyen, Dat and Alam, Firoj et.al [27] present a image filtering meth-

ods of during crises using deep learning techniques and perceptual hashing, the methods mainly conduct filtering and collect the important information of social media images.In the case of disasters and lack of labeled data, Li Xukun and Caragea Doina et.al [28] proposed a domain-adaptive countermeasure neural network method to recognise disaster images and detect damaged areas. Meng, Lingchao and Wen, Kuo-Hsun [29] to verify the correlation between the physical health of the elderly and the urban space, the Baidu Street View (BSV) of Macau Peninsula is used as the research scene, and the deep learning technology is used to perceive the high-density urban street space. Kim, Dongeun and Kang, Youngok et.al [30] proposed a understanding tourists' urban images with geotagged photos using convolutional neural networks.With the continuous increase of the urban population, the human gathering area has gradually evolved into a local dense temporal and spatial dynamic distribution. In order to better understand the urban environment, Chen, Meixu and Arribas-Bel et al [31] constructed an advanced image recognition model and used The marked Flickr pictures are used to train the neural network to quantify the feature information of different cities. Jayasuriya, Maleen and Arukgoda, Janindu et.al [32] present a novelty localising PMDs perception methods of urban street via convolutional neural networks. the method combines two important components, one of which uses 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 CNN to detect common environmental landmarks such as tree trunks for positioning. However, to further enhance the perceptions ability of human for urban and villages form , environmental and infrastructure, Wang, Tianyi and Tao, Yudong et.al [33] present a new multi-task and multi-modal deep learning framework with automatic loss weighting to assess damage after disaster events. Agarwal, Mansi and Leekha et.al [34] proposed a towards multi-modal damage analysis methods to reply deployment, challenges and assessment and are called Crisis-DIAS. In order to promote human response to disaster events and extract as much detailed information as possible from limited data, Alam, Firoj and Ofli, Ferda et al [35] collected a large amount of multi-modal data (include image and text) from Twitter, effectively solving the limitation of lack of labeled image data, and improving The ability to respond to and manage disasters.

Summary, although the above methods use deep learning technology to improve people's perception of 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.), which ultimately leads to large perceptual errors. Secondly, these methods do not take into account the changes in the same area and different time periods.

3. Our proposed methods

In the section, we will elaborate on our proposed spatiotemporal perceptions framework from three aspects: the feature extraction of urban and villages image, network structure of backbone and adjustment and optimization.

3.1. Overview

With the rapid development of society and economy, and field surveys of urban and rural residents' gathering places or other non-gathering 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 the changes in the gathering place. However, when external factors are more complex, if you use traditional machine learning methods to perceive changes in the same area at different periods of time, you are susceptible to interference from these external factors, such as light, occlusion, etc., 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, and improve the perception accuracy. Therefore, in order to detect the subtle changes in the urban and villages environment, form and infrastructure in different time periods 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 spatio-temporal sensing method to detect urban and rural changes from the perspective of remote sensing. The method mainly includes spatial branch and temporal branch. The temporal branch embeds the urban and villages 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 perceptions frameworks are shown in Figure 1. Considering that urban and rural images in the same area at different times have both relevance and differences in spatial and temporal,, we set the input images to $T^{(1)}$ and $T^{(2)}$ respectively,and $T^{(1)}, T^{(2)} \in R^{H \times W}$, where H, W indicates Height and width respectively, the image size of inputs is 1024×1024. The feature information of output via feature extraction module is $x^{(1)}, x^{(2)} \in R^{C \times H \times W}$, where C indicates the channel dimension. The spatiotemporal feature information are refine to attention feature map $y^{(1)}$ and $y^{(2)}$ via spatiotemporal perceptions module. However, the module are mainly composed of efficient channel attention guided squeeze-and-excitation. Then, we resizing the optimization feature information to the size of the input remote sensing images. Meanwhile, we will calculates the distance of each pixel pair in the corresponding feature maps and archive an corresponding distance map ζ in the proposed of optimization update.

3.2. Spatiotemporal feature extraction via DenseNet

In the past ten years, convolutional neural networks and improved convolutional neural networks have been widely used in urban and rural perception tasks relying on their strong learning ability, 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) and other dimensions to extract better detailed information. Compared with traditional hand-made or manual field survey methods, the method based on convolutional neural network not only has higher efficiency, but also shows stronger performance. In order to obtain better detailed information and different scales of spatiotemporal information, we introduce DenseNet to model urban and rural images in different phases, and use it as a feature extractor to capture multi-scale spatiotemporal information to further enhance perception.

Due to the large differences in the intentional connotation 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 inter-class or intra-class differences in a certain spatial dimension, and multi-scale DenseNet can highlight these differences through features such as feature multiplexing and information cross-layer connection, and 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 intentional connotations (including buildings, roads, etc.). Therefore, we removed the final fully connected layer and used different scales Densely connected blocks obtain multi-scale information of these intentional targets. DenseNet's high-level information is semantically accurate, but it cannot effectively determine the position of the intended target, that is, 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 fused the high-order and low-level layers in the spatial dimension to generate more refined representations. Quantify and evaluate the intentional goals of cities and villages from different angles. It is worth noting that both temporal branch and spatial branch use multi-scale DenseNet as the feature extractor. The extraction process of spatiotemporal features can be expressed as.

$$\begin{cases} x_l^{(1)} = H_l([x_0^{(1)}, x_1^{(1)}, \ldots, x_{l-1}^{(1)}]) \\ x_l^{(1)} \in R^{C \times H \times W} \\ x_l^{(2)} = H_l([x_0^{(2)}, x_1^{(2)}, \ldots, x_{l-1}^{(2)}]) \\ x_l^{(2)} \in R^{C \times H \times W} \end{cases} \tag{1}$$

Where, l indicates the number of layers and $l \geq 1$. $H(\bullet)$ indicates the operate of DenseNet. $x_0^{(1)}$=$T^{(1)}$, $x_0^{(2)}$=$T^{(2)}$.

3.3. Spatiotemporal perceptions with cross-channel interaction attention

In order to further perceive the changes in the intentional connotation of cities and villages in recent years,

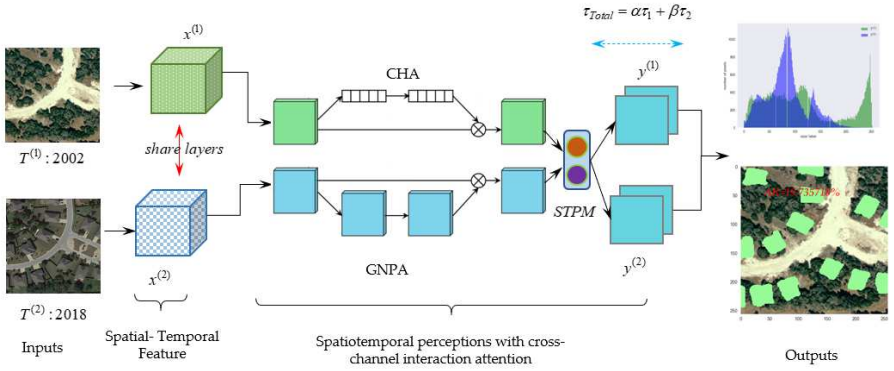

Fig. 1. The network structure of our proposed perceptions frameworks. Where $STPM$ indicates the module of spatiotemporal perceptions. $T^{(1)}$ and $T^{(2)}$ indicates the remote sensing image of different time phases. $x^{(1,2)}$ indicates the spatiotemporal feature information via feature extraction module, the module mainly contains densely connected convolutional networks. τ_{Total} indicates the total loss of our frameworks. τ_2 and τ_1 indicates the loss of spatial and temporal respectively. $y^{(1,2)}$ indicates the output feature via $STPM$. α and β indicates an learnable weighting factor.

and 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 designed an SE-enhanced channel attention The force module captures the rich global spatio-temporal relationships among the intentional individuals throughout the entire time and space, and establishes effective long short-term dependencies, so as to highlight the perception of subtle changes and temporal and spatial characteristics, and provide subsequent urban and rural planning, disaster evaluation, and intention information statistics. Provide reliable theoretical support for other tasks.The specific intention perception can be divided into the following steps:

Step 1. we first fused the captured multi-branch spatiotemporal information and denoted as.

$$\begin{cases} x_{Temporal} = Conv_{1\times1}(x_l^{(1)}, \cdots, x_l^{(1)}) \\ x_{Spatial} = Conv_{1\times1}(x_l^{(2)}, \cdots, x_l^{(2)}) \end{cases} \tag{2}$$

Where $x_{Temporal}$ and $x_{Spatial}$ indicates the multi-branch spatiotemporal information. $Conv_{1\times1}$ indicates the operate of convolutional and the kernel size is 1×1,

Step 2. we would $x_{Temporal}$ divided into γ subspace along the channel of feature maps and are denoted as $x_{Temporal} = [x_{Temporal_1}, \cdots, x_{Temporal_\gamma}]$. Then, the specific temporal semantics information of urban and villages intended object via each subspace $x_{Temporal_i} \in R^{\frac{C}{\gamma}\times H\times W}$ and generate a corresponding coefficient. The structure of the spatial branch is similar to the temporal branch, namely, can indicates as $x_{Spatial_i} \in R^{\frac{C}{\varepsilon}\times H\times W}$ and ε is subspace.

Step 3. In order to make the module more portable and more conducive to the statistics of global information, we use the spatiotemporal information of 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 villages intended object via cross-channel attention component are shown as.

$$x_{\gamma1}^{CH} = \delta(W_1 \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_{Temporal_{\gamma1}}(i,j) + b_1) \cdot x_{Temporal_{\gamma1}} \tag{3}$$

However, the feature information captured by the spatial branch is used as the input of the Group-Norm position attention (GNPA) 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 GNPA component can are obtained with.

$$x_{\varepsilon1}^{GNPA} = \delta(W_1' \cdot GN(x_{Spatial_{\varepsilon1}}) + b_1') \cdot x_{Spatial_{\varepsilon1}} \tag{4}$$

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

Meanwhile, to ensure efficiency and reliability, and help form 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.

$$W_\gamma = \begin{bmatrix} w^{1,1} & \cdots & w^{1,\gamma} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,\gamma+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-\gamma+1} & \cdots & w^{C,C} \end{bmatrix} \tag{5}$$

Where, w_γ indicates the weighting factor.

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

$$x_{STPM} = \left[x_{\gamma 1}^{CH}, x_{\varepsilon 1}^{GNPA}\right], x_{STPM} \in R^{C \times H \times W} \tag{6}$$

we using different branches to capture the characteristic information of urban and rural intentional targets can 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 urban and rural intentional targets. Interactivity, improve the ability of follow-up applications.

3.4. Optimization

To future improve the representations and perceptions ability of these spatiotemporal information for urban and villages intentions object changes, we present a loss functional of reconstruction. The loss functional are indicates as.

$$\tau_{Total} = \alpha \tau_{spatial} + \beta \tau_{temporal} \tag{7}$$

Where α and β is a learnable balance factor.

Summary ,we can better perceive changes in the content of urban and rural intentions in this way, achieve as much as possible the automated processing of the content of urban and rural intentions, and improve the ability and efficiency of emergency response after disasters.

4. Experiments discusses and analysis of spatiotemporal perceptions

In the sections, we would detial describe our perceptions results of urban and villages intention object, and provide a discusses and analysis

4.1. Data preparation and processing

In view of the fact that there is no database specifically used to perceive changes in urban and rural intentional targets, we have screened public baseline datasets such as LEVIR-CD and b.

LEVIR-CD: 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 concentrated on the growth of various types of buildings (Such as villas, high-rise apartments, small garages and large warehouses) in cities and villages.

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

However, in order to better perceive changes in urban and rural intentional goals and provide more reliable experimental support for subsequent urban planning, intention type or disaster evaluation, we have preprocessed these initial data to ensure that the processed data set is suitable for urban and rural areas. The description of intentional connotation is more comprehensive, and it is also more suitable for urban and rural perception tasks.

4.2. Training configuration

To achieve better perceptions effect of city and villages intention object via training our proposed frameworks. we conduct a sequence of initial setting for the frameworks and enhance these datasets by augmenatation methods. Meanwhile, the augmenatation can also be effectively to make up for the lack of urban and ru-

ral intention content data. such as rotation, noise, color change, etc. The processing of the datasets are shown as Figure 2.

For networks structure of our present spatiotemporal perceptions frameworks, the scale are set as $s \in \{S = 1, 2, 3\}$, the learning rate are set as $1e - 4$. the $Dropout$ is 0.5, the epoch is set as 600. However, to further ensure the effectively of training for our frameworks, We force the input remote sensing image size to be compressed to 256×256. The datasets are divide into three subsets: training (40%), testing (60%) and Validation (10%).

4.3. Experimental describe

To demonstrating effectively of our proposed spatiotemporal perceptions frameworks, which also helps to collect information on the connotation of urban and rural intentions, and improve the responsiveness of tasks such as urban planning, disaster evaluation and intention type judgment. We have tested and verified on two data such as LEVIR-CD and SZAB, with Precision, Recall and F-score as an evaluation indicator. Meanwhile, we will give the change area of the intention content in the urban and rural images, namely, AR. According to Table 1, we can draw the following conclusions, Our proposed spatio-temporal perception framework has achieved the best performance on the two public baseline datasets, namely, the F-scor is 88.04% and 53.72% respectively. The main reason is that the perception framework we designed uses multi-branch deep neural networks to first capture the deep semantics and shallow physical appearance information of urban and rural intentional targets, and describe the intentional targets from different levels and scales. Secondly, in order to further establish Spatio-temporal dependence and 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 intentional connotation of urban and rural areas. ability. In addition, we can also find that only using DenseNet (Our(No-STPM)) for spatio-temporal information extraction can also achieve better performance, but compared to using the STPM module (Our), its F-score value is reduce 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 social economy, the urban and rural forms will also undergo earth-shaking 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, and at the same time, it is more accurate, providing a certain experimental basis for subsequent urban planning and assumptions.

In order 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 as in Figure 3.

4.4. Ablation studies

In order to further verify the influence of different components on the proposed framework, experimental tests are carried out based on LEVIR-CD datasets, and relevant perception results and analysis are given. The perceptions results are shown in Table 2.

According to Table 2, we can find that the perception accuracy of using CHA (our(No-GNPA)) is obviously better than using GNPA (Ours(No-CHA)) , and 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 may be that CHA captures more effective specific information and is more sensitive to urban and rural intentional object. However, in order to better show the impact of CHA and GNPA components on the overall frame performance, we have given a visual hotmap of different components. The hotmap are denoted in Figure 4.

According to Figure 4, we can obviously seed that the two components are used in conjunction to form information complementarity, which can better express 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 urban and rural intentional targets, while GNPA can better locate the target's location, and their collaborative work can establish a more effective dependence.

4.5. The discusses of results

In order to show that the proposed perception framework can effectively detect the intentional connotations of urban and rural environments, forms, and infrastructure, and contribute to various tasks such as

Fig. 2. The augmenatation results of LEVIR-CD and SZAB baseline datasets. (a) indicates the original image of urban and villages. (b) and (c) indicates they are rotated by 15 degrees and 150 degrees respectively. (d) indicates gaussian noise is added.

Table 1
The perceptions results of our proposed frameworks. Where AR indicates the percentage before and after the average area change.

Datasets	Model	Precision(P%)	Recall (R%)	F-score(F%)	AP(areas %)
LEVIR-CD	ours(Non-STPM)	80.15	88.94	80.51	27.13
	ours	84.38	92.04	88.04	31.43
SZAB	ours(Non-STPM)	43.22	63.31	51.37	12.46
	ours	46.15	64.27	53.72	13.49

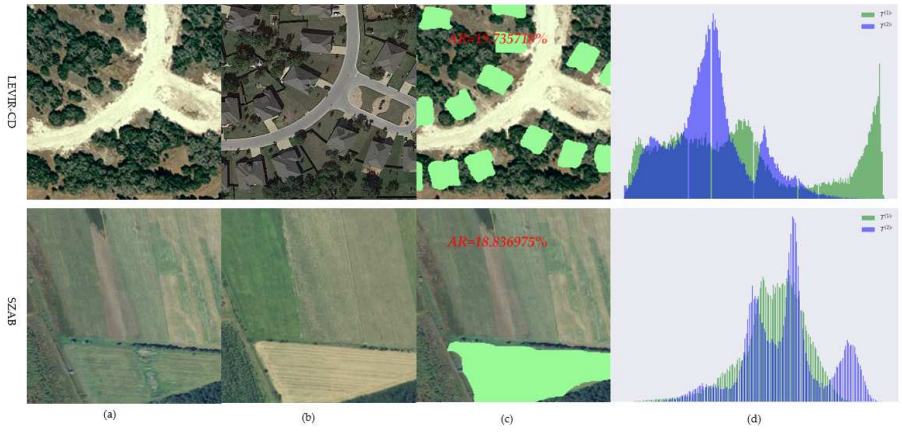

Fig. 3. The perceptions results of our present frameworks. (a) and (b) indicates the image urban and rural areas before and after the change. (c) indicates the perceptions results, where the white represents the part of the perceived change, the AR indicate the areas of change. (d) indicates the histogram.

disaster evaluation and intention type statistics, we show the perception results of multiple intentional targets. The result is shown in Figure 5.

5. Conclusions and next studies

In the paper, we to perceive the changes in the connotation of urban and rural intentions, present a ex-

Table 2
Frequency of Special Characters

Model	Precision(P%)	Recall (R%)	F-score(F%)	AP(areas %)
Ours(No-CHA)	81.54	89.35	85.27	28.09
Ours(No-GNPA)	82.97	90.69	86.66	28.88
Ours	84.38	92.04	88.04	31.43

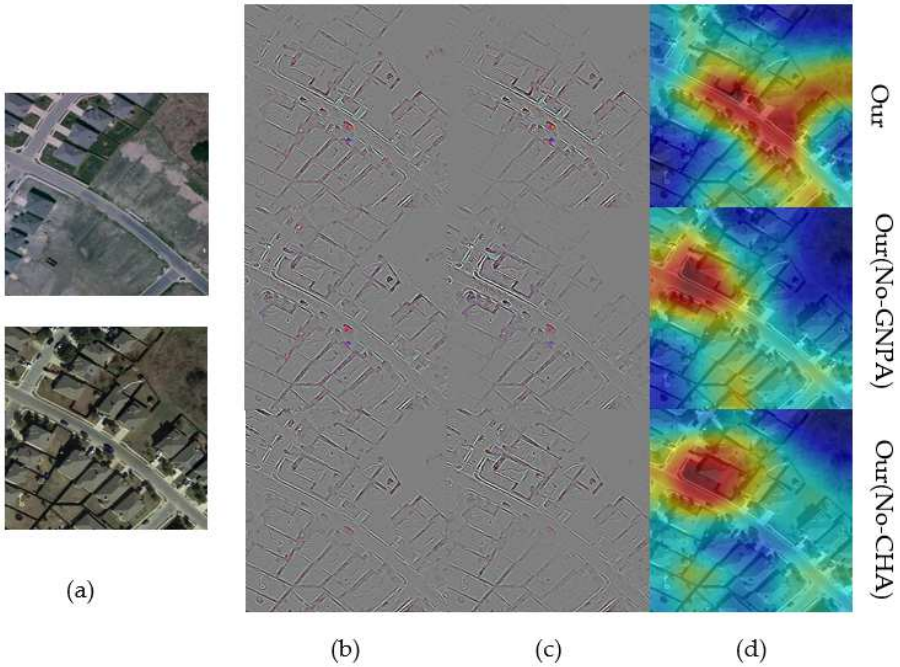

Fig. 4. The hotmaps of different components. (a) indicates the initial image of urban and rural areas before and after the change. (b) and (c) indicates the spatiotemporal feature maps of middle layers. (d) indicates different components hotmap.

ploring spatiotemporal change of city and village from remote sensing using multi-branch networks.The perception framework not only effectively captures the multi-scale spatiotemporal information of the intended target, but also uses STPM to capture the long-term spatiotemporal correlation, describing the intended target 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 these characteristic information are 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 good performance.

Although the framework has achieved good perceptual performance, the results of intentional targets with large scale changes are poor and need to be improved. Next, we will develop a simpler and more effective semantic framework to guide deep neural networks to large scale changes. The urban and rural intentions of the target are better characterized.

Declarations

Availability of data and materials
The datasets used and analysed during the current study are available from the corresponding author on reasonable request.
Competing interests
The authors declare that they have no competing interests.
Funding
This study was funded by National Natural Science Foundation of China—Youth Science Fund (grant number:51308427.Funding is mainly used for data processing and analysis in my manuscript.

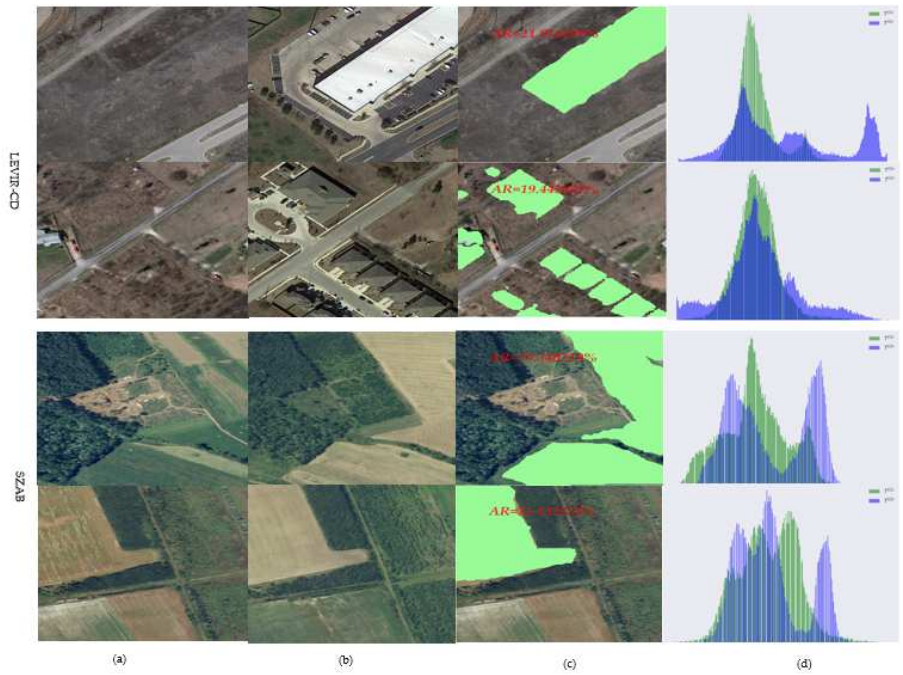

Fig. 5. The perceptions results of dour proposed frameworks. (a) and (b) indicates the image urban and rural areas before and after the change.(c) indicates the perceptions results, where the white part of (c) represents the perceived change part. (d) indicates the histogram of urban and villages intention object

Authors' contributions

MZ:Conceptualization, Methodology,Software:Programming, implementation of the computer code and supporting algorithms ,Writing- Original draft preparation.

YT:Writing-Review Editing,Supervision,Funding acquisition.

Acknowledgements

Not applicable.

References

[1] Naik N, Philipoom J, Raskar R, et al. Streetscore-predicting the perceived safety of one million streetscapes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014: 779-785.

[2] Gebru T, Krause J, Wang Y, et al. Using deep learning and google street view to estimate the demographic makeup of the us[J]. arXiv preprint arXiv:1702.06683, 2017.

[3] Li X, Cai B Y, Ratti C. Using street-level images and deep learning for urban landscape studies[J]. Landscape Architecture Frontiers, 2018, 6(2): 20-30.

[4] Wegner J D, Branson S, Hall D, et al. Cataloging public objects using aerial and street-level images-urban trees[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 6014-6023.

[5] Zhou H, Liu L, Lan M, et al. Using Google Street View imagery to capture micro built environment characteristics in drug places, compared with street robbery[J]. Computers, Environment and Urban Systems, 2021, 88: 101631.

[6] Maniat M. Deep Learning-Based Visual Crack Detection Using Google Street View Images[D]. The University of Memphis, 2019.

[7] Obeso A M, Benois-Pineau J, Vázquez M S G, et al. Saliency-based selection of visual content for deep convolutional neural networks[J]. Multimedia Tools and Applications, 2019, 78(8): 9553-9576.

[8] Morbidoni C, Pierdicca R, Paolanti M, et al. Learning from Synthetic Point Cloud Data for Historical Buildings Semantic Segmentation[J]. Journal on Computing and Cultural Heritage (JOCCH), 2020, 13(4): 1-16.

[9] Hu Y, Manikonda L, Kambhampati S. What we instagram: A first analysis of instagram photo content and user types[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2014, 8(1).

[10] Hochman N, Manovich L. Zooming into an Instagram City: Reading the local through social media[J]. First Monday, 2013.

[11] Jett J, Senseney M, Palmer C L. Enhancing cultural heritage collections by supporting and analyzing participation in Flickr[J]. Proceedings of the American Society for Information Science and Technology, 2012, 49(1): 1-4.

[12] Liu Z J, Wang J, Liu W P. Building extraction from high resolution imagery based on multi-scale object oriented classification and probabilistic Hough transform[C]//Proceedings. 2005 IEEE International Geoscience and Remote Sensing Sympo-

sium, 2005. IGARSS'05. IEEE, 2005, 4: 2250-2253.

[13] Li J, Xie X, Zhao B, et al. Identification of Urban Functional Area by Using Multisource Geographic Data: A Case Study of Zhengzhou, China[J]. Complexity, 2021, 2021.

[14] Berg A C, Grabler F, Malik J. Parsing images of architectural scenes[C]//2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007: 1-8.

[15] Bose A, Chowdhury I R. 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[J]. Modeling Earth Systems and Environment, 2020, 6(4): 2235-2249.

[16] Liu X, Tian Y, Zhang X, et al. Identification of urban functional regions in chengdu based on taxi trajectory time series data[J]. ISPRS International Journal of Geo-Information, 2020, 9(3): 158.

[17] Mathias M, Martinovic A, Weissenberg J, et al. Automatic architectural style recognition[J]. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2011, 3816: 171-176.

[18] Chu W T, Tsai M H. Visual pattern discovery for architecture image classification and product image search[C]//Proceedings of the 2nd ACM International Conference on Multimedia Retrieval. 2012: 1-8.

[19] Tardioli G, Kerrigan R, Oates M, et al. Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach[J]. Building and Environment, 2018, 140: 90-106.

[20] Gadal S, Ouerghemmi W. Identification of urban objects using spectral library combined with airborne hyperspectral imaging[C]//4ème colloque du Groupe Hyperspectral de la Société Française de Photogrammétrie et Télédétection (SFPT-GH). 2016.

[21] Manzoni M, Monti-Guarnieri A, Molinari M E. Joint exploitation of spaceborne SAR images and GIS techniques for urban coherent change detection[J]. Remote Sensing of Environment, 2021, 253: 112152.

[22] Llamas J, M Lerones P, Medina R, et al. Classification of architectural heritage images using deep learning techniques[J]. Applied Sciences, 2017, 7(10): 992.

[23] Chen Q, Cheng Q, Wang J, et al. Identification and Evaluation of Urban Construction Waste with VHR Remote Sensing Using Multi-Feature Analysis and a Hierarchical Segmentation Method[J]. Remote Sensing, 2021, 13(1): 158.

[24] Attari N, Ofli F, Awad M, et al. Nazr-cnn: Fine-grained classification of uav imagery for damage assessment[C]//2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2017: 50-59.

[25] Vetrivel A, Gerke M, Kerle N, et al. 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[J]. ISPRS journal of photogrammetry and remote sensing, 2018, 140: 45-59.

[26] Hamdi Z M, Brandmeier M, Straub C. Forest damage assessment using deep learning on high resolution remote sensing data[J]. Remote Sensing, 2019, 11(17): 1976.

[27] Nguyen D T, Alam F, Ofli F, et al. Automatic image filtering on social networks using deep learning and perceptual hashing during crises[J]. arXiv preprint arXiv:1704.02602, 2017.

[28] Li X, Caragea D, Caragea C, et al. Identifying disaster damage images using a domain adaptation approach[C]//Proceedings of the 16th International Conference on Information Systems for Crisis Response And Management. 2019.

[29] Meng L, Wen K H, Zeng Z, et al. 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[J]. Sustainability, 2020, 12(5): 1799.

[30] Kim D, Kang Y, Park Y, et al. Understanding tourists' urban images with geotagged photos using convolutional neural networks[J]. Spatial Information Research, 2020, 28(2): 241-255.

[31] Chen M, Arribas-Bel D, Singleton A. Quantifying the Characteristics of the Local Urban Environment through Geotagged Flickr Photographs and Image Recognition[J]. ISPRS International Journal of Geo-Information, 2020, 9(4): 264.

[32] Jayasuriya M, Arukgoda J, Ranasinghe R, et al. Localising PMDs through CNN Based Perception of Urban Streets[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020: 6454-6460.

[33] Wang T, Tao Y, Chen S C, et al. Multi-task multimodal learning for disaster situation assessment[C]//2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2020: 209-212.

[34] Agarwal M, Leekha M, Sawhney R, et al. Crisis-dias: Towards multimodal damage analysis-deployment, challenges and assessment[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(01): 346-353.

[35] Alam F, Ofli F, Imran M. Crisismmd: Multimodal twitter datasets from natural disasters[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2018, 12(1).