IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 July 2024, accepted 3 August 2024, date of publication 12 August 2024, date of current version 23 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3441537

RESEARCH ARTICLE

Evaluation of Public Participation and Emotion Classification for the Reconstruction of Old Communities Based on Semantic Analysis Algorithm

MENGYA GAO
School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan 430070, China
e-mail: 245238@whut.edu.cn

ABSTRACT This paper delves into the assessment of public engagement and the categorization of emotional states to gain deeper insights into individuals' involvement in the revitalization of historical communities and their emotional experiences during this process. First, this paper proposes an enhanced Faster R-CNN model for detecting and analyzing public participation. Specifically, our model incorporates attention mechanisms, both in the spatial and feature channels, to enhance the identification of human subjects and facilitate more precise target localization. Additionally, this paper addresses the challenge of detecting multi-scale targets by employing Feature Pyramid Network (FPN) technology, which enriches the features computed by Faster R-CNN. Furthermore, a method for describing public emotions based on semantic analysis is proposed. Leveraging semantic analysis algorithms, double LSTM and self-attention mechanisms are employed to categorize the emotions of individuals and generate captions to understand the emotions. Experiments validate the effectiveness of our approach in accurately pinpointing individuals within images, enabling a comprehensive assessment of public participation in the restoration of historic residential areas. Our method obtains a CIDEr value of 127.0 and an F value of 0.8856. Moreover, our method proficiently characterizes participants' emotional states, providing valuable technical support for societal development.

INDEX TERMS Image captioning, improved faster-RCNN, public participation, reconstruction of old communities, transformer.

I. INTRODUCTION

The revitalization of historic residential areas constitutes a multifaceted endeavor, entailing intricate interconnections among material space, society, economy, and even ethical considerations. At its core, this process entails the realignment of interests, intricately intertwined with the fundamental well-being of every resident residing within. Public participation stands as an indispensable facet of urban renewal, holding paramount importance for both the decision-making processes of revitalization and the enduring development of communities. The extent and profoundness of public involvement bear profound significance for the deliberations surrounding reconstruction and the sustainable evolution of neighborhoods.

With the improvement of modern society, more old residential areas need to be renovated and updated to match the new lifestyle. As the owners, they are always needed to provide valuable advice in the renovation of old communities. Presently, the theoretical and practical exploration of public participation in the revitalization of historic residential areas primarily centers on the advancement of public involvement and participation models. Regrettably, scant attention has been devoted to the evaluation methodologies and emotional states associated with public participation. Arnstein [1] delineates the extent of public engagement into three

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno M. Garcia.

tiers: dominant participation, symbolic participation, and non-participation, further encompassing eight distinct stages. In our nation, public participation can be categorized into three modes: ''top-down,'' ''bottom-up,'' and a fusion of both. The ''top-down'' approach endows governmental entities with decision-making authority, introducing planning proposals to the public and soliciting input from stakeholders in a hierarchical manner. Conversely, the ''bottom-up'' approach places the public in a central role throughout the planning process, from defining planning objectives to crafting planning schemes, with minimal government intervention. The amalgamation of these approaches entails a collaborative decision-making process wherein both the government and the public share equal negotiation rights in shaping the overall planning trajectory.

To evaluate the owners' participation during the reconstruction process of the old communities and ensure that the requirements of the owners for the reconstruction of the residential area are met, this paper carried out a study on the evaluation method of public participation and the classification of emotional states [2]. The evaluation task of public participation is transformed into the problem of detection and positioning of owners during the rebuilding process of the old communities, and the classification of public emotion is quantified into the research of a public emotion description method based on semantic analysis.

Deep learning provides technical support for crowd detection. The accuracy rate of the detection algorithm in dense scenes provides a certain guarantee for the security of the scene. However, the overlap rate between people in dense environments is too high, so it brings certain challenges for the algorithm to accurately detect pedestrians with high overlap rates. At present, most object detection systems are constructed based on the proposed framework. Among them, the classic networks include YOLO [3], [4], [5], [6] which realizes an end-to-end single-phase detection algorithm, and Faster-Rcnn [7], [8], [9], [10] which includes RPN. These algorithms have achieved good results in conventional target detection applications. Among them, the YOLO believes that detection is the regression problem, uses a single network to complete target classification and location, and avoids the stage of candidate box extraction, realizing high real-time detection. For the first time, Selective Search used in the Fast-RCNN algorithm is replaced by the RPN network to realize end-to-end training. However, when detecting highly overlapped objects, it is difficult for the detector to generate differentiated predictions for each suggestion box, and serious overlap will inevitably lead to incorrect Suppression of NMS (Non-Maximum suppression).To solve this problem, some researchers have tried to use a new Loss function (Aggregation Loss), complex non-maximum suppression (Softer NMS), and the addition of an FPN network to Faster-RCNN to solve the problem of multi-scale detection [11], [12]. However, using the improved network for case detection with a low overlap rate will lead to performance degradation.

Despite their objective portrayal of community reconstruction participation, existing detection methods fall short in accurately capturing the semantic essence of community scenes. Consequently, we rely on language descriptions to gain a comprehensive understanding of scene content. In the generation phase of natural language text descriptions, we devise an image-to-text escape model that incorporates pre-extracted image features and a recursive neural network (RNN). To enrich the RNN's input processing capabilities, we integrate Visual Attention, multi-language models, and other algorithms. This ensures that input features embody a sense of image hierarchy, three-dimensionality, subjects, and predicates, thereby furnishing robust features for the image escape process. Our primary research endeavors encompass: 1) Establishing an RNN model that incorporates weighted image semantic features as input, allowing for the propagation of previous states and convolutional layer features into the RNN at each iteration, enabling significant semantic elements to permeate into the final natural language text description. 2) Facilitating RNN-based language text generation, leveraging features and models from preceding steps to generate captions for spatial images, realizing an end-to-end transformation from image to text. In the analysis module, we refine existing natural language text descriptions by incorporating multi-language grammar models, scene semantic information directed graphs, and other algorithms. This streamlines text descriptions, differentiates the conciseness of described targets, and maps the intricate relationships between multiple targets, providing solid foundations for the escape of text descriptions.

The objective of this research paper is to meticulously assess the engagement level and emotional wellbeing of the public amidst the revitalization of aging communities. This assessment aims to be swift and precise, leveraging the capabilities of detecting, locating, and visually describing individuals through sensor imagery. Therefore, this paper proposes a public participation detection method based on improved Faster-Rcnn and a public emotion description method based on semantic analysis, to accelerate the replacement of the old residential areas. The main contributions are as follows:

1) This paper proposes a novel public people detection approach aimed at identifying individuals' involvement in the community reconstruction process. 2) This paper proposes a public emotion description methodology that integrates semantic information to gain insight into the emotional states of individuals participating in community reconstruction.

II. RELATED WORKS

This section will introduce some research status related to this paper. Considering that the reconstruction of old residential areas is related to people, to obtain the information related to people accurately, the research status of pedestrian detection methods is introduced first. The participation of community reconstruction requires a deep understanding of human behavior and expression with words. Therefore, this paper

introduces the research status of image description methods. At present, the methods of person detection are mostly used in the congestion detection problem of small-scale sparse scenes. In this method, each pedestrian instance is accurately detected by the detection algorithm, and then the congestion detection is implemented according to the number of instances. This method based on pedestrian detection is not fundamentally different from the general target detection tasks, most of which are directly transferred to the target detection tasks. However, the current crowd congestion detection task often needs to detect a large number of people in many open scenes. The pedestrian detection-based method is not competent for the increasingly complex congestion detection task, so the research and use of such methods are relatively few in the research field. Aiming at the problem of crowd congestion detection under extremely crowded conditions, Wang et al. [13] propose a depth model on the network structure of AlexNet [14]. This model uses the strategy of expanding negative samples in the training set (the number of people is zero) to alleviate the complex background interference problem. Shang et al. [15] input the entire image into the pre-trained Google LeNet model to obtain high-level semantic features and then build a mapping relationship between these features and the number of people in the local area. The biggest advantage of this algorithm is the integration of both global and local context information.

Previous methods (non-deep learning), such as Baby Talk [16] Every Picture Tells a story [17], etc., apply the corresponding operations to achieve image features. After SVM classification and so on, the objects of an image can be obtained. With objects and attributes, CRF is applied to recover the captions of the pair of images. This approach relies heavily on 1) the extraction of image features and 2) the rules needed to generate sentences, which is naturally not ideal. So far, the latest algorithm research points [18], [19], [20], [21] are all focused on structural adjustment and function increase of CNN or RNN [22], [23], [24].

III. THE METHOD FOR ECALUATION OF PUBLIC PARTICIPATION AND EMOTION CLASSIFICATION DURING THE RECONSTRUCTION OF OLD COMMUNITIES

A. IMAGE PERSONNEL POSITIONING AND SEMANTIC ANALYSIS

In the semantic analysis of spatial images, it is often difficult to summarize the information in the image into one feature information due to the complexity of the number, type, and location of the objects in the image. This study intends to first use deep learning to detect and recognize various objects in the image as far as possible, then use the information on the convolutional layer of the image to calculate the significance weight, to reflect the relationship between the objects, and finally obtain the analysis results by integrating each object and various relationships contained in it.

There are usually obvious differences between objects in spatial images. However, in many cases, the position of objects in the image overlaps, which is easy to cause misdetection or missing detection of object type recognition. To solve this problem, this study intends to carry out in-depth research on target recognition algorithms (TR) in traditional deep learning, to improve its breakthrough in recognition rate and accuracy. The main ideas are as follows:

(a) To design an appropriate deep learning neural network architecture for mobile terminals, it should have a relatively small memory ratio, moderate computing time, and a deep enough network depth. This network architecture will be specifically developed in this study and serve as the research basis;

(b) Using the designed network architecture, an algorithm model is constructed to solve the problem of missing detection and wrong detection. It is planned to first carry out undifferentiated detection on all targets in the image, obtain their position coordinates and intercept corresponding image areas, and then design a target classifier for these intercepted areas to obtain its corresponding target type, thus forming a cascade algorithm architecture, as shown in Figure 1. This algorithm architecture is to be developed and improved in this study.

FIGURE 1. Schematic diagram of Cascading architecture of Faster R-CNN and CNN.

After the targets in the image are identified, they need to be fused into a feature for subsequent use. Therefore, CNN features will be used to reflect the interaction between targets. After the convolution of each object in the image, it does not have a response in all the convolution channels but appears in the specific channels. These specific channels are extracted and then processed by spatial saliency to get the final feature map. This study intends to establish a weighted method of target significance in each layer of CNN to achieve the establishment of appropriate relationships between targets. The main research methods are as follows:

Set the weights among different convolutional layers of CNN used. Considering that the information contained in CNN's shallow layer is not enough to represent the essential relationship between objects, according to this condition, the processing only focuses on the deep convolutional layer or pooling layer, and the weight plan of each layer follows the following formula:

$$T_l = \phi\chi T_{l-1}\ (T_0 = 1) \tag{1}$$

where l refers to the convolutional layer of the l -th layer, and Tl denotes the weight of the l -th convolution layer. χ and ϕ will be formulated later.

(b) Calculate the significance of the object across the different convolution layers of CNN. In the process of weight calculation, it is divided into two parts: one is to calculate the weight χi,j of the current convolution layer space (i, j

represent the width and height of the graph, and m and n are the coordinates of points on the feature graph). The calculation method is as follows:

$$\chi_{i,j} = \frac{S'_{i,j}}{\left(\sum_{m,n} S'_{m,n}{}^2\right)^{\frac{1}{2}}} \tag{2}$$

where S' is the feature matrix obtained after the current convolution layer is accumulated on the depth, and the feature map with depth k is defined as Fk, so, as shown in the formula:

$$S' = \sum_k F^k \tag{3}$$

The second is to calculate the weighting ϕk at the depth of the current convolution layer, the calculation method is shown in the pseudo-formula:

$$\phi_k = \begin{cases} \log\left(\frac{\sum_x N_x}{N_k}\right), & N_k > 0 \\ 0, & N_k = 0 \end{cases} \tag{4}$$

where Nx is calculated as shown in the following formula:

$$N_x = \sum_{i,j} 1[\lambda^{i,j,x} > 0] \tag{5}$$

which means the sum of the quantities of all samples is calculated while the conditions are met.

As shown in Figure 2, because of the significance weights that have been obtained, they need to be applied to the corresponding convolution layer and to generate the final image semantic features. FPN constructs a pyramid structure by leveraging convolutional feature maps from different levels, enabling the network to extract features across various scales. This approach facilitates the detection and segmentation of objects of diverse sizes. By integrating low-level, high-resolution feature maps with top-level, low-resolution ones, FPN achieves scale invariance, allowing the network to detect or segment objects independent of their size within the image. Furthermore, FPN bridges the gap between feature maps of different scales and the detection head (e.g., anchor frames or bounding box regressors), thereby assisting the detector in locating targets at multiple scales and enhancing person detection capabilities. This study intends to use the softmax function, which is widely employed, and other methods to process the weighted feature maps. The calculated weight Tl is used to the feature map Ai,j,k of the l-th layer. The weighted operation is carried out and the final image semantic feature O is obtained, as shown in the following formulas:

$$O' = A_{i,j,k} \prod_0^d F_l \tag{6}$$

$$O = \text{soft}\max\left(O'\right) \tag{7}$$

where d is the depth of the deep learning model and softmax is the normalized exponential function.

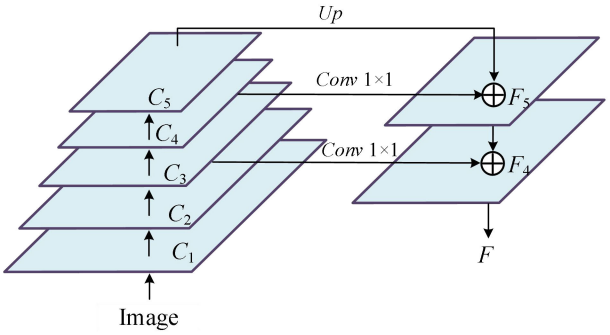

FIGURE 2. The structure of FPN.

B. GENERATION AND ANALYSIS OF PERSONNEEL IMAGE LANGUAGE DESCRIPTION

For the human brain, the most direct understanding of a spatial image is to describe it in words. Therefore, this study intends to convert the semantic features of previously generated images into the most intuitive language description and classify and analyze the obtained text description, to make adequate preparation for the subsequent research. This process can be divided into two parts:

Due to the mechanism of CNN, the image semantic features previously generated have fixed lengths, which will reduce the accuracy of image translation for long sentences. To solve this problem, this study intends to add an image content discrimination mechanism (CDM) to judge whether the image content has numerous targets and complex relationships. Then, using the previous salience mechanism, the targets in different locations correspond to a feature and provide the image semantic features of the whole image and features containing location information for RNN decoding. The main research ideas are as follows:

(a) Establish the image content discrimination mechanism. The main purpose of image content discrimination is to estimate whether there are many relationships among the objects in the image, but it cannot be completely screened. For this purpose, this study intends to use the combination of the number of objects and the number of types to make a judgment, which needs to be confirmed in specific studies.

(b) Establish an image semantic feature decoding model based on the salience mechanism. Up to now, all the work done is the processing of image features, and this part is the transformation of image features into text descriptions. For images with different levels of complexity, the processing methods are also distinct. For simple images, the semantic features of the whole image are taken as model input and decoded using RNN. The following formula is derived:

$$x_{-1} = CNN\,(I) \tag{8}$$

$$x_t = W_e S_t, t \in \{0 \ldots N-1\} \tag{9}$$

$$p_{t+1} = RNN\,(x_t), t \in \{0 \ldots N-1\} \tag{10}$$

where I represents the input image, and the correct literal description of the image is $S = (S_0, S_1, \ldots, S_N)$. S_0 is the

initial word, SN describes the end word, We is the model embedding the word, pt+1 is the probability and x is correct at the next moment of time t. Then, its Loss is calculated as follows:

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t) \tag{11}$$

where pt refers to the probability distribution of the generated word.

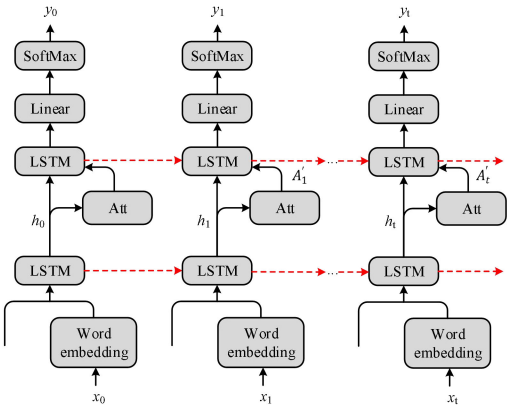

FIGURE 3. The structure of RNN.

For images with complex content, the research is still carried out based on RNN, as shown in Figure 3. As the relationship between objects is complex, two key quantities need to be paid attention to. One associates with time t and corresponds to decoding time. Another one is the target ai of the input sequence, which can correspond to a regional position of the image. The way to do this is to calculate a weight ati for each object i in the input sequence at time t with Softmax. Two aspects need to be included: one is the region ai being calculated, and the other is the information ht−1 at the previous time t-1, the derivation formula is shown as follows:

$$e_{ti} = f_{att}(a_i, h_{t-1}) \tag{12}$$

$$a_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})} \tag{13}$$

where fatt is the scoring function that computes the two information of target i and time t coupled. Then, a certain mechanism κ will calculate ai and ati. This mechanism is to be developed in detail in this study.

Dual LSTM is capable of capturing both forward and backward contextual information of the input data simultaneously. Dual LSTM ensures that the model considers not only the current part of the image but also the relevant information before and after it when making emotional classification decisions. When faced with complex emotional states, the attention mechanism can help models focus on key information related to emotions. The attention mechanism can dynamically adjust the focus of attention based on the characteristics of the input data. When dealing with complex emotional states, models may need to simultaneously attend to multiple factors related to emotions, and the attention mechanism can dynamically allocate attention weights based on the importance of these factors, enabling a more comprehensive analysis of emotional states. By focusing on key information related to emotions, dynamically adjusting the focus of attention, and combining contextual information, models can comprehensively understand emotional tendencies.

C. ANALYSIS OF EMOTIONAL TEXT DESCRIPTION OF PERSONNEL

In actual situations, it is often necessary to use this generated text description conversion to generate images, to achieve the purpose of rapid transmission. This study intends to analyze the text description in this part. First, it is divided into simple sentences and long difficult sentences, and different sentence patterns are analyzed in different ways, to make full preparation for the subsequent image generation. The main research ideas are as follows:

(a) To establish a text description classification model. The main purpose of the classification of text description is to predict whether the grammar or content of the text description is complex or whether there are too many accompanying states for a particular object in the description. For this purpose, this study intends to establish a classification model by synthesizing the number of objects, the number of types, and the complexity of the relationship between objects in the text description. The establishment of this model will be carried out in the specific research.

(b) To build a long and difficult sentence description analysis model. For the description of simple sentences classified by the former, this step is skipped. This part analyzes the description of long difficult sentences. For the processing of the description of long difficult sentences, the most critical thing is the category of the target in the text description and its relationship. In this study, the Scene Graph is intended to be used to process the description of long difficult sentences. In Scene Graph, nodes refer to objects and edges can be the interaction between objects. Take the English description as an example, as shown in Figure 4, so it can represent the whole content of a description. Suppose a set C of preset object categories, and its set of related categories is R, then a Scene Graph can be represented as where O is the target set in the description, as shown in the formula:

$$O = \{o_1, \ldots, o_n\}, o_i \in C \tag{14}$$

Then, its corresponding relation set E, as shown in the formula:

$$E \subseteq O \times R \times O \tag{15}$$

The edges between the concrete objects are shown as follows:

$$(o_i, r, o_j)\, o_i, o_j \in Or \in R \tag{16}$$

With the Scene Graph, each object and relationship in the literal description is quantified as an embedded vector that serves as input data to the image generation model.

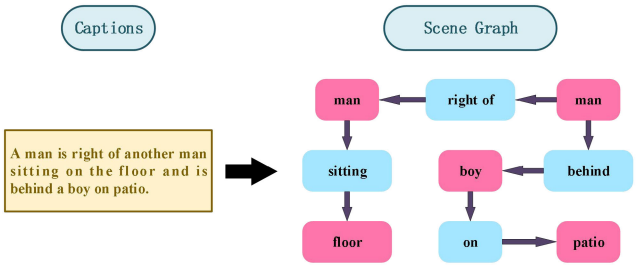

FIGURE 4. Structure principle of scene graph.

IV. EXPERIMENT AND ANALYSIS

A. DATASET AND IMPLEMENT DETAILS

This paper use 3 d Street View dataset (https://github.com/amir32002/3D_Street_View) to evaluate the performance of our model. The dataset consists of Street View data with camera pose, a 3D model of 8 cities, and extended metadata. The data set is huge, with the Street View dataset alone containing 25 million images and 118 million matching image pairs.

The experiment was carried out on the device with i7-12700 Cpu and Rtx 3090 GPU, the operating system was Unbuntu, and the network model was implemented under the Pytorch framework. The experiment was configured with a total of 1000 rounds, a batch size of 16, and an initial learning rate set at 0.1. Following a single round of warm-up, a learning rate attenuation strategy was implemented, with an attenuation rate of 0.8. The model utilized SGD as its optimizer, featuring a momentum of 0.9, and a weight decay value set to 1×10^{-4}.

B. RESULTS AND DISCUSSION

To evaluate the model performance, BLEU [25], METEOR [26], ROUGE-L [27], and CIDEr [28] are used as evaluation indexes to measure the model performance. BLEU index represents the n-tuple correlation among the description statements and the manually marked reference statements and represents the accuracy rate between the two sentences. METEOR uses WordNet to match specific sequences, which gives it a stronger correlation with manual evaluation; The ROUGE-L index is a measure of the co-occurrence accuracy and recall rate, which reflects the co-occurrence probability of the common clause of the generated statement and the reference statement, and reflects the adequacy and fidelity of the generated statement. CIDEr index regards each sentence as a document and represents it in the form of TF-IDF (term frequency-inverse document frequency) vector, and obtains cosine similarity between the generated statement and reference statement by weight calculation of each n-tuple. This index reflects the consistency of the image content information.

Experiments were carried out on the 3D Street View dataset and compared with previous advanced technologies: Baby-talk, Swin-Transformer [29], and Transformer [30]. The results are shown in Table 1. Our model obtained the best human emotion description performance and detection and positioning accuracy, and the segmentation accuracy reached 88.65%(DSC). Compared with the Swin-Transformer, the evaluation index of Cider is improved by about 1.33%. As shown in Table 2, after comparison with several mainstream networks at present, compared with other networks that introduce attention mechanisms, such as Swin-Transformer and Transformer. Our model has fewer parameters, faster reasoning times, lower computational complexity, and is more lightweight. The pure CNN-based approach is less sensitive to emotional information and prone to over-interpretation problems. Although the method based on pure Transformer is sensitive to semantic emotional information, it is prone to insufficient description due to the loss of some features. For example, Swin-Transformer's description of emotional states is missing, while our model describes them correctly and retains good boundary information. Experimental results show that compared with Swin-Transformer and Transformer, as well as other pure CNN-based frameworks, our model pays more attention to semantic information and can achieve better emotional prediction. For the pure Transformer approach, our model both ensures sensitivity to emotional information and prevents feature loss.

TABLE 1. Comparison with other methods.

	Baby-talk	Transformer	Swin	Ours
B@1	77.2		77.4	78.4
B@2	-		61.5	62.5
B@3	-		47.6	48.7
B@4	36.2	34.3	36.5	37.8
M	27.0	26.4	27.4	29.1
R	56.4	55.2	56.8	58.3
C	113.5	196.1	114.4	127.0
S	20.3	19.7	20.5	22.23

In addition, the methods Yolo v4, Deit [31], SSD, and Vit [32] are choosen to conduct a comparative experiment with our model because of the test results of public participation during the reconstruction of old communities.

As shown in Table 3, it can demonstrate that our method surpasses other methods in all indicators, especially F-value, the comprehensive evaluation index of detection and positioning. Our method is far superior to other methods. FPN extracts multi-level feature information, information fusion improves feature interaction ability, aggregates all global features for the deep fusion of deep and shallow features, and finally improves the accuracy of model detectors.

The experimental results clearly demonstrate that our model strikes a balance between sensitivity to emotional information and prevention of feature loss. This is achieved through the effective combination of FPN and our

TABLE 2. The results of parameters, cost time, and GFLOPs of the methods.

	GFLOPS	Cost time(ms)	Parameters
Baby-talk	43.95	185	32.3
Transformer	39.36	174	84.5
Swin	42.69	198	99.4
Ours	35.48	193	63.8

TABLE 3. The results of parameters and accuracy of the recognition methods.

	Recall	Precision	F	Parameters
Yolo v4	0.8336	0.9023	0.8755	123.6
Deit	0.8296	0.8956	0.8562	95.3
SSD	0.8248	0.8759	0.8487	86.4
Vit	0.8387	0.8972	0.8714	96.7
Ours	0.8484	0.8996	0.8856	75.6

custom-built deep learning architecture, which allows the model to focus on semantic information while preserving crucial details. Overall, our approach represents a significant step forward in the field of individual recognition and emotion analysis in complex, real-world environments.

Finally, this paper conducts ablation experiments to evaluate each module.This experiment conducts in-depth and independent measurements of the performance of TR and CDM in Table 4. Based on the Baseline model, TR and CDM are introduced separately to observe their respective impacts on the model's performance comprehensively. The experimental data indicates that when TR and CDM are individually integrated into the Baseline model, the model's various evaluation metrics show significant improvements. Specifically, the addition of TR leads to a 1.6 increase in CIDEr. Similarly, the incorporation of CDM also brings remarkable performance gains to the model. Furthermore, embedding CDM into the Baseline model simultaneously is able to further elevate CIDEr to 127.0, which is not just a simple superposition of their individual effects. This suggests that there exists a mutually enhancing relationship between TR and CDM, enabling them to jointly optimize the model's predictive capabilities.

C. VISUALIZATION

To illustrate the performance of the public engagement assessment method and sentiment classification, this paper chooses four images and show the results of our method on these images.

Firstly, with regards to public engagement, our personnel detection methodology adeptly computes the number of individuals contributing to the revitalization of the old residential area portrayed in the figure. Specifically, as exemplified

TABLE 4. Ablation experiments.

TR	CDM	B4	M	R	C
Baseline		26.4	20.3	54.2	124.2
O		27.8	21.3	56.9	125.8
	O	28.0	21.2	56.5	126.1
O	O	29.1	22.2	58.3	127.0

in Figure 6, our system precisely identifies every individual within the frame (denoted by blue bounding boxes). Notably, though, in scenarios where individuals are heavily obscured, our approach exhibits minor variations or fluctuations in detection accuracy. Nonetheless, aside from addressing occlusion challenges, our methodology shines in handling objects of diverse sizes, attributable to the integration of Feature Pyramid Networks (FPN). Furthermore, in the realm of public emotion classification, our model swiftly captures and categorizes the emotional state of the public through the generation of descriptive narratives. As evident from the red-highlighted text in the sentences showcased, each crafted sentence poignantly captures the emotional atmosphere surrounding the public's participation in the renovation of the old communities, encompassing sentiments such as 'seriousness' among others, thereby enabling real-time monitoring of public sentiment.

In addition, we use a project example to demonstrate the effectiveness of this method, the general layout is shown in Figure 5. Geguang Community is located in Hongshan District, Wuhan, with a total land area of 40,308 square meters and a total of 18 buildings and 920 households. The community was built before 2000. In order to complete the renovation task of the old community that needs to be rebuilt before 2000, the renovation project of the community was launched in early 2022. The renovation process of old residential areas includes nine stages: policy publicity, information collection, willingness survey, site survey, program preparation, joint review, program publicity, parallel approval, design, etc. In Figure 7, the planner and designer conducted a survey on the renovation intention of residents in the community, and our model also generated scene statements captions, which can demonstrate our model's performance.

To sum up, the research on the evaluation method of public participation and the classification of emotional states carried out by us can be fully used during the rebuilding of old communities, to ensure that the reconstruction can be recognized by the public.

The experiment proves that the performance of the semantic analysis algorithm based on the old community reconstruction public participation evaluation and emotion classification method has reached the advanced level. This method, combined with semantic analysis algorithms, can more accurately assess the degree of public participation in

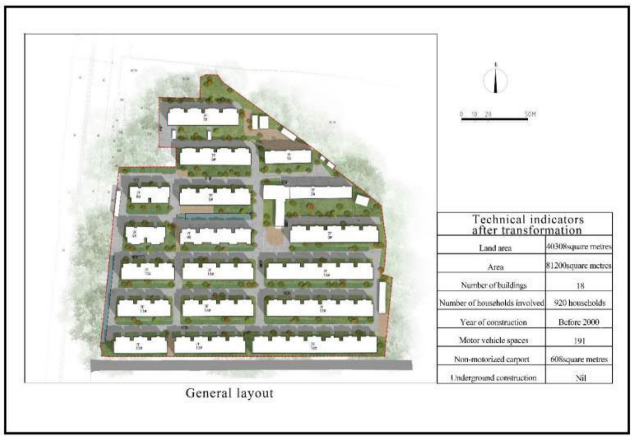

FIGURE 5. General layout of Geguang community.

FIGURE 6. The visualization of our method.

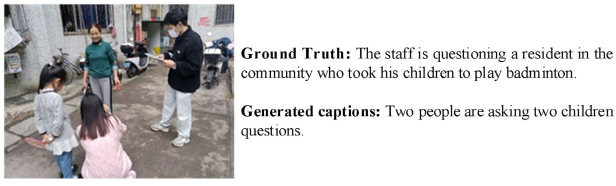

FIGURE 7. The visualization of the project.

the reconstruction of old neighborhoods. Whereas traditional methods may be based only on qualitative analysis or simple statistics, this approach allows for a deeper understanding of the level of public involvement in a redevelopment project. Through emotion classification, the method can identify and record the different emotional states of the public in the reconstruction process of the old community. This helps decision makers to better understand the public's emotional experience in order to improve project planning and management in a more targeted manner. Methods Using semantic analysis and deep learning techniques, quantitative data support is provided to enable decision makers to make decisions based on detailed evaluation and sentiment classification results. This kind of data support is more reliable than traditional subjective assessments. By gaining insight into the public's level of engagement and emotional state, this approach can provide powerful guidance for community building. This helps ensure that the project meets public needs and increases community satisfaction and acceptance.

V. CONCLUSION

This paper proposes a public participation detection method based on improved Faster-Rcnn and a public emotion description method based on semantic analysis to find out how people took part in rebuilding old communities and how they felt at the time. The attentional mechanism and FPN technology are used to make the human target in the image stand out, which makes it easier for the detector to find the target. Then, according to the detected human target, the emotion of the human is described by a semantic analysis algorithm based on the transformer and self-attention mechanisms to realize the classification study of public emotion. The experimental results show that our model can accurately locate the people in the picture, to evaluate the public's participation in the reconstruction of the old communities. In addition, our method can also accurately describe the emotional state of the participants and realize their emotional classification, providing technical support for the construction of society.

DATA AVAILABILITY

The dataset employed in this investigation is made readily available and accessible to interested parties.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

ACKNOWLEDGMENT

The author would like to thank the anonymous reviewers whose comments and suggestions helped to improve the manuscript.

REFERENCES

[1] S. Arnstein, "A ladder of citizen participation," *J. Amer. Inst. Planners*, vol. 35, no. 4, pp. 290–302, 2020.
[2] B. Chakravarthi, S.-C. Ng, M. R. Ezilarasan, and M.-F. Leung, "EEG-based emotion recognition using hybrid CNN and LSTM classification," *Frontiers Comput. Neurosci.*, vol. 16, Oct. 2022, Art. no. 1019776.
[3] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Proc. Comput. Sci.*, vol. 199, pp. 1066–1073, Jan. 2022.
[4] M. J. Shafiee, B. Chywl, F. Li, and A. Wong, "Fast YOLO: A fast you only look once system for real-time embedded object detection in video," 2017, *arXiv:1709.05943*.
[5] W. Lan, J. Dang, Y. Wang, and S. Wang, "Pedestrian detection based on YOLO network model," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Aug. 2018, pp. 1547–1551.
[6] M. Hussain, "YOLOv1 to v8: Unveiling each variant—A comprehensive review of YOLO," *IEEE Access*, vol. 12, pp. 42816–42833, 2024.
[7] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[8] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang, "Revisiting RCNN: On awakening the classification power of faster RCNN," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 453–468.
[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.
[10] F. Xin, H. Zhang, and H. Pan, "Hybrid dilated multilayer faster RCNN for object detection," *Vis. Comput.*, vol. 40, no. 1, pp. 393–406, Jan. 2024.
[11] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proc. IEEE/CVFConf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6459–6468.
[12] R. Tang, Z. Liu, Y. Song, G. Duan, and J. Tan, "Hierarchical multi-scale network for cross-scale visual defect detection," *J. Intell. Manuf.*, vol. 35, no. 3, pp. 1141–1157, Mar. 2024.
[13] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1299–1302.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.
[15] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1215–1219.
[16] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "BabyTalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.
[17] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, and J. Hockenmaier, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2010, pp. 15–29.
[18] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2015, pp. 2048–2057.
[19] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6298–6306.
[20] A. Mathews, L. Xie, and X. He, "SemStyle: Learning to generate stylised image captions using unaligned text," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8591–8600.
[21] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5561–5570.
[22] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.
[23] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1179–1195.
[24] X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, F. Huang, and R. Ji, "RSTNet: Captioning with adaptive attention on visual and non-visual words," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15460–15469.
[25] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, Assoc. Comput. Linguistics*, Philadelphia, PA, USA, 2002, pp. 311–318.
[26] B. Satanjeev, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Evaluation Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
[27] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Association Computational Linguistics*, Barcelona, Spain, 2004, pp. 74–81.
[28] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[31] H. Touvron, M. Cord, M. Douze, F. Massa, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

MENGYA GAO received the B.S. degree in environmental art design from Southeast University, in 2010, and the M.S. degree in fine arts school from Shandong University, in 2014. She is currently pursuing the Ph.D. degree in civil engineering and architecture with Wuhan University of Technology. Her research interests include theory and methods for the reconstruction of old communities, landscape architecture planning, and protecting of built heritage.