SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2022, Vol. 65 112104:1–112104:11
https://doi.org/10.1007/s11432-019-2899-9

AR-CNN: an attention ranking network for learning urban perception

Zhetao LI[1,2], Ziwen CHEN[1,2], Wei-Shi ZHENG[3*], Sangyoon OH[4] & Kien NGUYEN[5]

[1]*Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan* 411105*, China;*
[2]*Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education, Xiangtan University, Xiangtan* 411105*, China;*
[3]*School of Data and Computer Science, Sun Yat-sen University, Guangzhou* 510006*, China;*
[4]*Department of Computer and Information Engineering, Ajou University, Suwon* 443-749*, South Korea;*
[5]*Graduate School of Engineering, Chiba University, Chiba* 263-8522*, Japan*

Received 25 October 2019/Revised 20 January 2020/Accepted 27 April 2020/Published online 24 December 2021

Abstract An increasing number of deep learning methods is being applied to quantify the perception of urban environments, study the relationship between urban appearance and resident safety, and improve urban appearance. Most advanced methods extract image feature representations from street-level images through conventional visual computation algorithms or deep convolutional neural networks and then directly predict the results using features. Unfortunately, these methods take color and texture information together during processing. Color and texture are prime image features, and they affect human perception and judgment differently. We argue that color and texture should be operated differently; therefore, we formulate an end-to-end learning methodology to process input images according to color and texture information before inputting it into the neural network. The processed images and the original image constitute three input streams for the triad attention ranking convolutional neural network (AR-CNN) model proposed in this study. In accordance with the aspects of color and texture, an improved attention mechanism in the convolution layer is proposed. Our objective is to obtain the scores of humans on urban appearance in accordance with the prediction results computed from pairwise comparisons generated by the AR-CNN model.

Keywords ranking network, urban perception, attribute learning, attention network, colour and texture

Citation Li Z T, Chen Z W, Zheng W-S, et al. AR-CNN: an attention ranking network for learning urban perception. Sci China Inf Sci, 2022, 65(1): 112104, https://doi.org/10.1007/s11432-019-2899-9

1 Introduction

With the rapid economic growth and development, people have become more attentive to their surroundings, and the physical appearance of cities has directly affected people's sensory judgment. Urban appearance is important because it affects the behavior and physical and mental health of city residents and the ideas held by policy makers. Improving urban appearance is beneficial to future city development. According to broken windows theory [1] and the evidence found by social scientists, poor urban appearance deteriorates urban environments in terms of social security issues, education problems, and population loss, eventually leading to urban recession. Many countries worldwide have already implemented policies to improve the urban appearance; such policies include the "Quality of Life Program 2020" in Saudi Arabia and the "Quality of Life Program" of New York City.

In recent years, these studies have been limited to a few neighborhoods. Scientists have conducted field surveys or crowd-sourced research but have not fully utilized the global stereoscope image database. For example, Hong Kong has 4768 streets, each of which has images of streetscapes that can reach tens of thousands in number, making studies lack quantitative data on urban perception. At present, an increasing number of researchers have begun using convolutional neural networks to solve tasks that humans have difficulty accomplishing. This trend is due to the rapid growth in urban appearance data,

* Corresponding author (email: zhwshi@mail.sysu.edu.cn)

the excellent performance of deep learning, and the large amount of manually labeled data, such as the Place Pulse 1.0 (PP1.0) [2] and 2.0 (PP2.0) datasets [3], which are obtained through crowdsourcing. Researchers have used computer vision algorithms that are trained on street-level datasets to automate the processing of survey data on cities around the world.

Naik et al. [4] used the PP1.0 dataset combined with a common image feature descriptor to convert perceived security into a fraction of the perceived response to the image; such response is a classic representation of street scores. Dubey et al. [3] proposed the ranking Streetscore-CNN (RSS-CNN) algorithm, which uses the PP2.0 dataset. These datasets consist of data collected through a crowd-sourced question-based game. In the game, one of two street-view images is selected in terms of which place looks safer, livelier, wealthier, mundane, exquisite, and dismal. The datasets contain 1.17 million pairs of comparison results, involving 110988 images in 56 cities on six continents. Compared with previous methods, the RSS-CNN algorithm introduces the Siamese convolutional neural network into perceptual pairwise comparison prediction, which only considers extracting the feature representation through generic convolutional neural networks. In the process of human perception, image colors and textures affect the human brain in different ways; however, previous studies do not quantify color and texture information individually. In other words, RSS-CNN only extracts the high-level features of images, thereby combining color and texture for processing.

Motivated by human perception, we propose a novel method called attention ranking convolutional neural network (AR-CNN) in this study. The AR-CNN algorithm solves the problem in which the effect of low-level image features on human perception is not processed appropriately (e.g., the color and texture information of the images). Our method inherits the ability of RSS-CNN to extract features, although an attention mechanism is added in the classification part to enhance the performance of the ranking task between the triplet networks. Such a model enables the individual processing of color and texture information; thus, their special effect for learning urban perception can be achieved. Lastly, our model combines the color and texture of images into high-level convolutional features. We use the PP2.0 dataset to train the convolutional neural network model, which is used to predict the pairwise comparisons of perceptual attributes. The experiments show that the accuracy of our proposed model is better than that of the RSS-CNN model.

The remainder of this paper is structured as follows. We present the analysis of urban perception and human perception in Section 2. In Section 3, we propose a novel method for predicting the result of the comparison pairs on the PP2.0 dataset. Then, the experiments and analyses are presented in Section 4. Lastly, we use the pretrained model to predict and analyze all images on the PP2.0 dataset.

The contributions of this work are summarized as follows:

(1) We propose a novel framework to predict the results of pairwise image comparisons for quantifying the perception of the urban environment.

(2) We use improved attention methods to maximize the separate quantification of the color and texture of images.

(3) We discard many fully connected layers and prune our model, thus improving our model's efficiency. As a result, the number of the RSS-CNN model's weight parameters is 263 million, whereas that of our method is 43 million at most.

2 Related work

This section discusses the relevant literature from two aspects: (1) human visual perception from the aspects of color, texture, and scene attributes and categories, and (2) prediction of perceived responses to urban streetscape images.

The literature on the prediction of perceptual responses of people to images is increasing daily. A major representative is aesthetics [5,6]. For example, Ren et al. [5] predicted the perceptual response to aesthetics from the aspect of images for personalized problems. This research was followed by studies on memory; Isola et al. [7] predicted perceptual response to the memory aspect of images, and Quercia et al. [8] pointed out the human perception aspect of urban street-level images. The existing color, texture, and visual vocabulary of images have also been analyzed. Gibson [9], a famous psychologist, stated that the seemingly paradoxical assertion will be made that perception is not based on sensation. That is, it is not based on having sensations. But it is surely based on detection information. Studies on color psychology show the ways by which different colors affect us. Colors can even reveal personality

traits based on a person's favorite color. Meanwhile, image texture perception consists of understanding images [10–12]. Trussell et al. [13] created a mathematical model that describes the effect of texture on the perception of color and found that color, texture, and other fundamental factors are essential for our perception. Thus, the color and texture in images can affect our perception.

Streetscore arithmetic extracts generic image features and then performs support vector regression [14] on the image feature representation. This algorithm is trained on data consisting of street images of some American cities. Streetscore completes the prediction task of perceived security. Ordonez and Berg [15] contributed to the PP1.0 dataset and used three features, namely, Gist, Sift, and DeCAF, and then trained the features on PP1.0 in accordance with three perception attributes: safety, uniqueness, and wealth. Then, the comparison results were predicted. Given the rapid development of neural networks, Porzi et al. [16] used the convolutional network part of AlexNet to extract features, proposed a latent detector to help two classifications, and compared it with ranking SVM. Porzi et al. [16] also determined the distribution of medium visual elements in security. Recently, the RSS-CNN model was proposed by Dubey et al. [3], and they also brought us the PP2.0 dataset. The RSS-CNN model is a Siamese network structure. After feature extraction using the pretrained model, two major branches, namely, the ranking and fusion networks, are formed.

3 Approach

In this section, we introduce the proposed attention-based ranking network to predict the perceived attribute score of street-view imagery. We use the triplet network in Figure 1 [17] to provide an overview of the overall model. The input to the network is a randomly cropped image O from a full-sized street-view image. First, after the input O is processed, three different inputs are extracted into the convolutional and pooling layers to extract the feature representation. The feature extraction part uses the pretrained weights of AlexNet, PlacesNet, and VGGNet. Second, the network continues to extract deeper and more abstract feature representations through the convolutional layer containing the attention mechanism. Furthermore, the network predicts the results of pairwise comparisons through the fully connected layer, which generally constitutes a triplet network structure.

3.1 Processing the color and texture of images

The traditional machine learning method predicts results by classifying street scenes. This method only uses original street scenes. Many theories and studies have indicated that color can be considered an element of design. Color evokes strong emotional responses and thus affects human perception. Texture also plays a key role in design; it provides a sense of depth, distance, and emotions. For example, the squares labeled A and B in the same picture have the same shade of gray, but we did not believe that they were the same color either; the proof is available on the website of MIT Professor Edward H. Adleson. The important task is to break down image information into meaningful components and thus perceive the nature of the objects in different views. These arguments also reflect the importance of color and texture to human perception. Therefore, we propose a learning methodology that automatically processes the color and texture information of an image and obtains them fused by the RSS-CNN network.

Given a street-view image O, we need three inputs of the convolutional neural network: the clustered image, the image processing by using the algorithm for extracting the texture feature, and the original image O_a, O_b, and O. First, O_a is an image obtained by clustering the main colors of the image using the mean-shift clustering algorithm [18–20]. This process generates a color block for separating the gradients of the colors between the color centers to enhance color intensity distribution and highlight color information; this step is operation A in Figure 1. This clustering algorithm can average the color near class by the class center and then fuse color information into our model conveniently using the following attention method.

Second, O_b is the result of the street-view image O processed by the Gabor algorithm [21]; this step is operation B in Figure 1. The image generated in this step only highlights the texture features in the image, that is, highlighting the main buildings, cars, streets, and trees. The outline is of the visual information and its texture information. This algorithm can only reserve texture information completely, and it can promote combining texture information in our model by the following attention method.

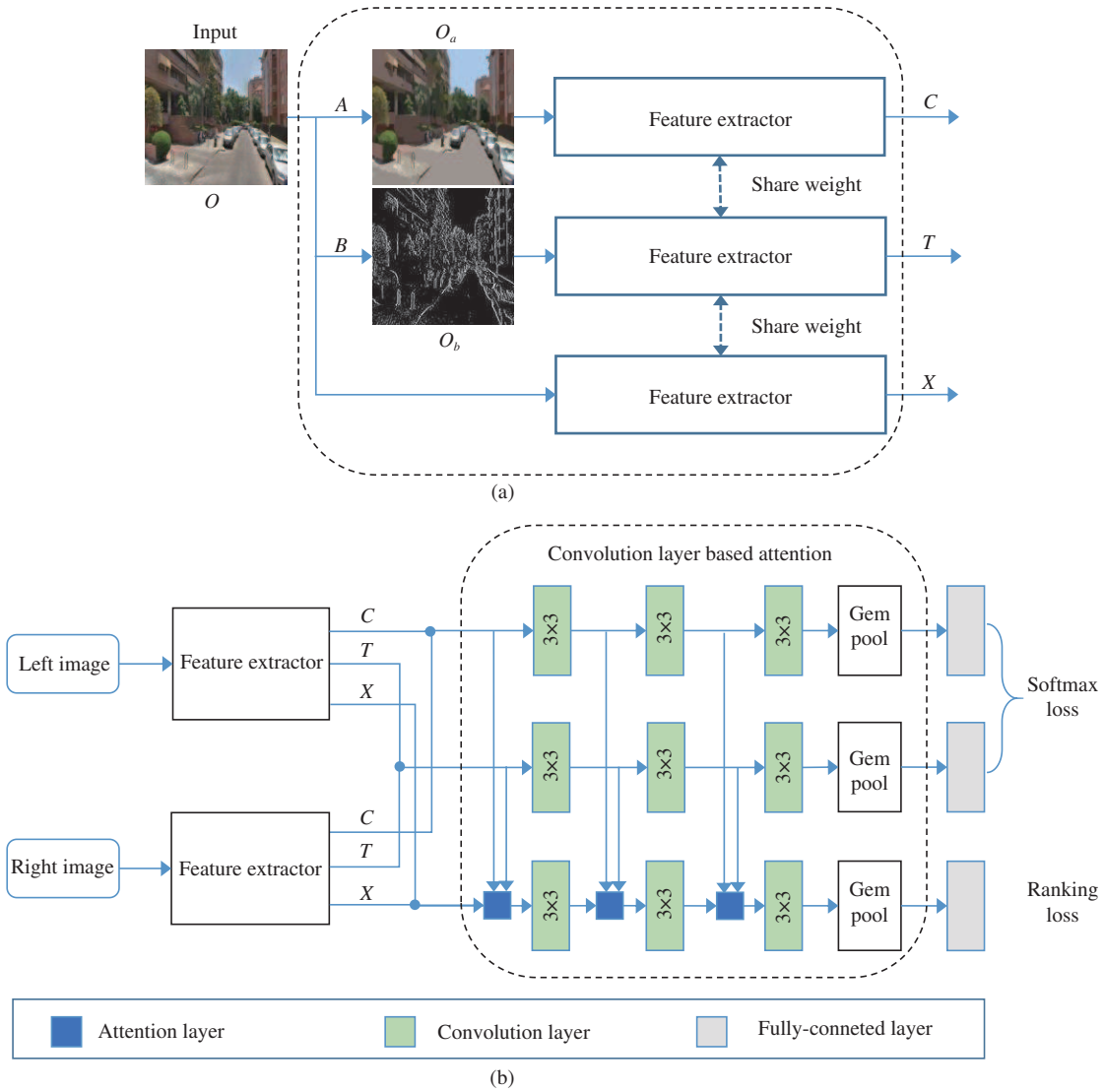

Figure 1 (a) Image preprocessing and feature extraction. (b) Convolutional layer based on the attention mechanism and output layer that consists of a generalized mean pooling [17] and a fully connected layer whose size is 1; the intersections of the feature extractors' output are connected through stacking. The feature extractor in (b) is the operation in (a). C, T, and X represent convolutional features obtained from the input images for the preprocessing of color, texture, and the original image, respectively. The figure shows the best CNN models (AR-CNN3), which have three attention layers in this model. This trained network has two losses, a ranking loss, and a softmax loss for predicting the pairwise comparisons of urban appearance. We experiment with AlexNet, PlacesNet, and VGGNet, which correspond to the feature extractor in this figure.

3.2 Fusing convolutional network by attention

Traditional neural networks based on attention mechanisms are generally built on recurrent neural networks, which are mainly used in the field of natural language processing. Moreover, RSS-CNN proposed by Dubey et al. [3] is based on the structures of the AlexNet, PlaceNet, and VGGNet networks, whose inputs are the original street scene images. Dubey fine-tuned these networks and added the fusion subnetwork to classify and predict the corresponding score of perceptual attributes, and the RSS-CNN achieved good results. Therefore, we use the structure of the three-layer convolutional layer in RSS-CNN to obtain high-level feature representation. Thus, we can extract the generalized feature of the three inputs by using the three branch networks shown in Figure 1. On this basis, we use the characteristic of the attention mechanism, which selectively weights some essential parts. For example, Yin et al. [22] applied the attention mechanism and convolutional layer to sentence modeling and proposed three methods for combining the attention mechanism, all of which have certain performance improvements. In our work, we must consider the correlation and difference between low-level and high-level features. Therefore,

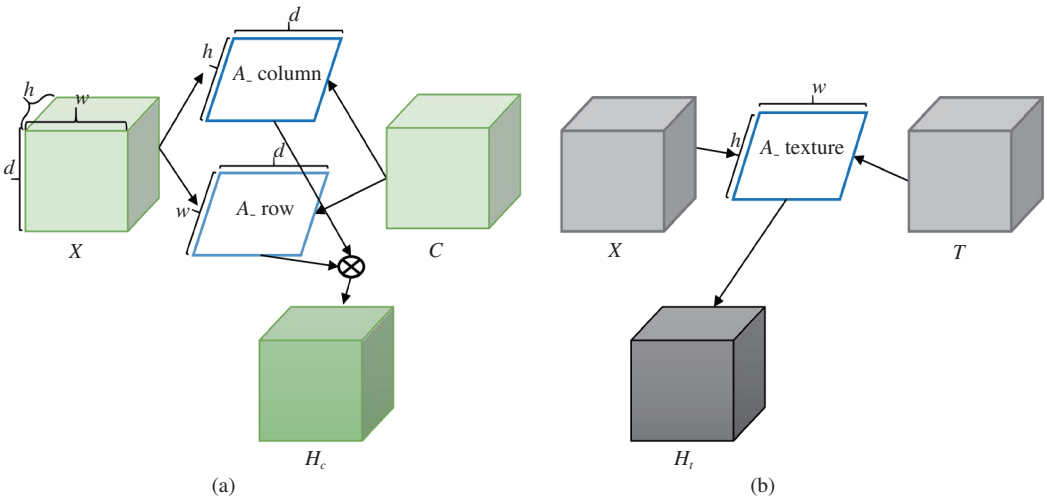

Figure 2 The operation of the attention layer consists of two parts: (a) computing the output of the colour attention and (b) computing the output of the texture attention. Finally, they will be fused to the final output of the attention layer.

we introduce the attention mechanism [22–29] into the convolutional layer of the latter half of the triad network structure in Figure 1.

Our objective is to integrate the two low-level features, namely, color and texture, into global features through an attention feature matrix while learning additional abstract features through a convolutional neural network. In Figure 1, the upper and middle branches of the network are used to extract local features of color and texture information, and the lower branches are used to extract the global features of images. Suppose the input of the attention mechanism structure shown in Figure 1 is X, C, and T, where X represents the output of the lower network branch structure of the triplet network in Figure 1; C represents the output of the upper network branch structure of the triplet network in Figure 1; T represents the output of the intermediate network branch structure of the triplet network in Figure 1; and X, C, and T are assumed to be three-dimensional feature vectors. The specific operation steps are as follows:

Color attention. The motivation is weighting the difference between feature X and feature C. On the basis of the attention mechanism calculation of the image color information, the information between the d feature maps of the feature vector C obtained by training the image processed in terms of color is considered transmitted to the feature vector X obtained from the original street-view image. The key to this step is calculating the column feature matrix $A_$column (A_c) and the row feature matrix $A_$row (A_r), as shown in Figure 2(a) (middle column). Here the attention feature matrix is intended to weight the importance of the present color in the images to the feature unit generated by the original image, thus perfectly fusing local features about color with global features.

The colour attention matrix is generated by matching the feature representation C and the feature unit of the feature representation X. The formula for calculating the A_r of the column feature matrix is defined as follows:

$$A_{r_{w,d}} = f\left(|X_{h,w,d}, C_{h,w,d}| \cdot \operatorname{sign}\left(X_{h,w,d}, C_{h,w,d}\right)\right), \tag{1}$$

where h represents the height of feature vector X, w represents the width of feature vector X, d represents the depth of feature vector X, $|\cdot|$ is the Euclidean distance, and the column feature matrix A_r is in the calculation of the Euclidean distance. The distance is calculated by accumulating the height h as the axis, and the sign(X, C) function is for calculating the symbol of the matrix element obtained by accumulating the difference between the input feature vector X and the input feature vector C with the height h as the axis. The multidimensional matrix is multiplied by the Euclidean distance point and then passed through the f function to obtain the column feature matrix A_r in Eq. (1). After experimental verification, the f function is $f(z) = (e^z - e^{-z})/(e^z + e^{-z})$. Similarly, with Eq. (1), we can calculate the row feature matrix A_c, which is different from (1) in accumulating the width(w) as the axis when calculating the row feature matrix A_c. Finally, the output H_c of the colour attention operation is obtained by the row feature matrix

A_r and the column feature matrix A_c, and the formula for the specific operation is defined as follows:

$$H_{c_{h,w,i}} = \left(A_{c_{h,1,i}} \times A_{r_{1,w,i}}\right) \cdot C_{h,w,i}, \tag{2}$$

where $i \in d$. Eq. (2) indicates that the row feature matrix A_r and the column feature matrix A_c are matrix multiplied by each element matrix corresponding to the depth d axis, and then the obtained attention feature weights are multiplied by the input feature C to obtain the output of the colour attention operation H_c.

Texture attention. The motivation is weighting the similarity of feature X and feature T. With the calculation of the attention mechanism of image texture information, the information between the d feature maps ($h \times w$) of feature vector T from the image processed in terms of texture is transferred to feature vector X from the original street scene image training, i.e., the attention feature matrix A shown in Figure 2(b) (middle column), where the attention feature matrix T is designed to weight the image. The importance of texture information to the feature units generated by the original image is also aimed at re-weighting the convolutional output of feature X, which perfectly fuses local features about texture into global features. A can be a new weight matrix, and the verification by Yin et al. [22] shows that it is feasible to combine this new feature mapping with the original feature representation as the input of the next convolutional layer. For this purpose, we define the formula for texture attention characteristics as follows:

$$A_{h,w} = f\left(\left|X_{h,w,d}, T_{h,w,d}\right| \cdot \operatorname{sign}\left(X_{h,w,d}, T_{h,w,d}\right)\right), \tag{3}$$

where the calculation of the f function and $|\cdot|$ is similar to that in 1. This step differs from the calculation of colour attention characteristics. The calculation of $A_{h,w}$ is based on the depth d axis;

$$H_t = A_{h,w} \times T_i, \tag{4}$$

where $i \in d$, which represents the attention feature $A_{h,w}$ and each element T_i in the input texture feature T, is multiplied by the matrix, and then the weight of the obtained attention feature is dot multiplied with the input feature C to obtain the output H_t of the texture attention operation. Since all of the above calculations gradually reduce the value of feature representation X, to avoid this situation, we define the final output as follows:

$$H = X + \gamma H_c + \beta H_t, \tag{5}$$

where γ and β are trainable hyperparameters. In our experiments, we initialize them with constants.

3.3 Classification and ranking

In this paper, we use the proposed convolutional neural network structure shown in Figure 1 to conduct both classification and ranking tasks. In other words, the classification loss and ranking loss are optimized simultaneously. The upper output and the intermediate output in Figure 1 are used to implement the classification task, which uses the softmax activation function. The lower output in Figure 1 is used to implement the ranking task, and the last activation function used is the tanh activation function. The purpose of using different loss functions is to optimize the loss of the network structure through different optimization approaches. Therefore, we define the classification loss function (L_c) as follows:

$$L_c = \sum_{(i,j,l)\in\{I,J,L\}} \sum_{k}^{K} -\hat{l}_k \times \log\left(f_k\left(x_i, x_j\right)\right), \tag{6}$$

where the triplet I, J, L is all images pairwise comparisons and $K = 2$; then, I represents the left street scene image, J represents the right street scene image, and L is the result of the comparison. The range of values is -1, 0, 1, that is, the right image is victory, they are equal, and the left image wins. We use two cases that do not contain equal in our experiments, and we transform the label into a one-hot vector in the classification task. $f(x_i, x_j)$ denotes the output of the upper or middle network in Figure 1 while computing classification loss.

The ranking loss function (L_r) is defined as follows:

$$L_r = \sum_{(i,j,l)\in\{I,J,L\}} \max\left\{0, \mu - (l \times p_r)\right\}, \tag{7}$$

Table 1 Results (%) of the prediction of the perceived attribute-safe

Approaches	AlexNet	PlacesNet	VGGNet
RSS-CNN	64.1	68.8	73.5
AR-CNN1 (ours)	63.2	67.9	71.9
AR-CNN2 (ours)	64.0	69.1	73.1
AR-CNN3 (ours)	65.3	70.2	74.6

where μ represents the hyperparameter, $p_r = f_1(x_i, x_j) - f_2(x_i, x_j)$, p_r denotes the difference between the output of image x_j on the left and that of image x_i on the right, but $f_k(x_i,x_j)$ represents the output of the underlying network in Figure 1 while computing the ranking loss. The ranking loss function is designed for our ranking task, which is consistent with the loss function formula of the RSS-CNN model. Finally, to train the convolutional neural network in Figure 1, we combine the classification loss function (L_c) in Eq. (6) and the ranking loss function (L_r) in Eq. (7) to train the total loss function: Loss $= L_r + \Theta_a L_c^a + \Theta_b L_c^b$, where L_c^a and L_c^b are the losses of the upper branch of the network and the media branch of the network, respectively; then, Θ_a and Θ_b are the coefficients that can weight the importance of the tasks. All hyperparameters are set by the grid search to obtain the maximum accuracy on the validation datasets.

4 Experimental

In this section, we will introduce the experiments and analysis from three aspects. First, we will accomplish two tasks, the classification task and the ranking task, through the convolutional network model shown in Figure 1; second, we analyze the correlation between the six perceptual attributes involved in the PP2.0 dataset through our model rankings. Finally, from the three aspects of the presence of colours, textures, and visual words in images, the distribution of the six attributes in these three aspects is systematically analyzed.

4.1 Predicting pairwise comparisons

We use AlexNet, PlacesNet, and VGGNet to initialize the parameter weights of the feature extraction part of AR-CNN. The number of attention network layers in AR-CNN is divided into the following: the AR-CNN model with one attention network (AR-CNN1), the AR-CNN model with two attention networks (AR-CNN2), and the AR-CNN model with three attention networks (AR-CNN3). In the experiment, the performance of image pairs is evaluated by subtracting the corresponding output of the lower-middle branch network in Figure 1 and then calculating the results using softmax. In this study, AR-CNN uses end-to-end learning on the basis of the combination of classification loss and ranking loss. AR-CNN3 achieves the highest accuracy (74.6%) for pairwise comparison prediction, and its performance surpasses that of the most advanced method in the world (RSS-CNN). We also compare color attention and texture attention; Using VGGNet and one attention mechanism, the accuracy rate is 71.4% when using only color attention for the safe attribute and 71.7% when using only texture attention.

In this subsection, pairwise image comparisons are classified, and the images are ranked. Some results are shown in Table 1. The average improvement in the predicted accuracy of all perceptual attributes is approximately 1%, and the results of two attributes, i.e., beautiful and wealthy, have better improvements of 71.8% and 67.4%, respectively.

4.2 Analyzing correlation between perceptual attributes

The dataset used in this study is PP2.0, which consists of six types of perceptual attributes: safe, lively, beautiful, wealthy, boring, and depressing. First, we use the neural network model proposed in this study to train all image pairwise comparisons corresponding to perceptual attributes. Then, we use Dubey's method with stable scores to generate random graphs. As input, image pairs generate numerous synthetic comparison results. Several comparisons are then input into ranking algorithms (such as TrueSkill [30]) to obtain stable ranking scores [3]. We use a pretrained AR-CNN3 (VGGNet) to generate pairwise comparisons of predictions, thus generating TrueSkill scores for six attributes. The 16 prediction results shown in Figure 3 are the images of six perceptual attributes from high to low scores. In this study, the TrueSkill score ranges from 0 to 10. When analyzing the correlation and independence between

Figure 3 Examples from the places pulse 2.0 dataset are ranked images from the result of the pairwise comparison generated by our proposed AR-CNN3 model.

different attributes, we use the TrueSkill score obtained by the aforementioned techniques to calculate the Spearman correlation coefficient in accordance with the score and then analyze the correlation among the six perceptual attributes. The pretrained AR-CNN3 model predicts the ranking scores of all images and includes uncovered images in this attribute; thus, this method is feasible. The results of the Spearman correlation coefficients are shown in Table 2. Table 2 shows that the correlation between the safe and

Table 2 Results of computing coefficients

R^2	Safe	Lively	Beautiful	Wealthy	Boring	Depressing
Safe	1.0	0.79	0.84	0.66	−0.35	−0.23
Lively	0.79	1.0	0.70	0.63	−0.72	−0.36
Beautiful	0.84	0.70	1.0	0.74	0.11	−0.27
Wealthy	0.66	0.63	0.74	1.0	0.19	−0.41
Boring	−0.35	−0.72	0.11	0.19	1.0	0.38
Depressing	−0.23	−0.36	−0.27	−0.41	0.38	1.0

Table 3 The average distribution (%) of the presence of colour in the images of the high scores ($\geqslant 8.0$) on the left and the low scores ($\leqslant 4.0$) on the right of all images in the PP2.0 dataset, and the units of the values in the table are percentages

Attribute	Red	Orange	Yellow	Green	Cyan-blue	Blue	Purple
Safe	10.3, 19.4	28.0, 57.6	20.0, 7.1	30.1, 0.8	2.0, 1.5	9.1, 13.5	0.5, 0.1
Lively	19.3, 1.7	23.0, 30.3	10.1, 6.3	10.3, 3.6	16.9, 0.1	20.2, 57.9	0.3, 0.0
Beautiful	1.3, 6.2	7.9, 33.0	9.5, 17.3	51.0, 2.2	5.2, 5.6	25.0, 30.1	0.1, 0.1
Wealthy	23.4, 9.1	21.5, 55.2	2.1, 13.7	11.3, 8.4	3.8, 2.6	56.2, 1.3	0.2, 0.1
Boring	0.1, 24.4	7.1, 23.8	13.3, 1.6	15.1, 8.9	0.4, 0.5	63.2, 30.2	0.1, 0.1
Depressing	9.6, 3.7	48.9, 32.9	6.7, 3.6	5.1, 12.6	3.7, 6.4	16.0, 40.7	0.0, 0.1

beautiful attributes is the highest. Second, the correlation between the safe and lively attributes is higher than that of the others, and the correlation between the beautiful and wealthy attributes is high. However, the safe attribute has a negative correlation with the boring and depressing attributes.

4.3 Analysing perceptual attributes

This section further analyzes how humans perceive images from the perspective of the basics of images or what elements of images affect the human perception of images. Through this analysis, we can also obtain some interesting and close to real-life phenomena. Here, we analyze all the images with high and low TrueSkill scores (high score is greater than or equal to 8.0, and low score is less than or equal to 2.0) generated from the prediction results of our proposed model. Moreover, this analysis contains the distribution of the presence of color, texture, and 'visual words' that are the scene attributes and the scene categories in the subsequent sections.

4.3.1 *Color*

To show how the color in the surrounding environment affects the human perception of the urban environment, we read the city street scene image according to the HSV color space, which facilitates the division of the HSV image into 10 types of common colors according to the range of pixels. However, environmental factors, such as time and weather, are not considered during image data acquisition; thus, we neglect black, gray, and white and only refer to seven colors (red, orange, yellow, green, cyan-blue, blue, and purple) to obtain the color distribution of each image. Then, we use the trick in Subsection 4.2 to generate the ranking scores of all street images contained in the dataset using the AR-CNN3 model that is pretrained on six different perception attributes and combined with the TrueSkill algorithm. On this basis, we combine the ranking scores of the attributes and the color distribution of each image and draw a conclusion through statistics, as shown in Table 3. This table shows that where people feel safer, the proportion of the two colors (green and yellow) is relatively higher than other colors. By contrast, where people feel more insecure, the proportion of the two colors (orange and red) is relatively higher than other colors. Where people feel more beautiful, the proportion of the color green is relatively high; conversely, the less beautiful humans feel, the higher the proportion of the colors blue and orange.

4.3.2 *Texture*

This subsection illustrates how to texture information in the image affects the human perception of the environment. For image texture analysis, only the basic texture information of the image is considered; such information includes regular (divided into edges, planes, and angles) and irregular types, in which we consider vertical, horizontal, 45-degree, and 135-degree edges. First, we transform the image into a texture matrix using the LBP algorithm and then classify the data points of the texture matrix in

Table 4 The average distribution (%) of the presence of texture in the images of the high scores ($\geqslant 8.0$) on the left and the low scores ($\leqslant 4.0$) on the right of all images in the PP2.0 dataset, and the units of the values in the table are percentages

Attribute	Edge	Flat	Corner
Safe	23.42, 20.17	25.88, 28.9	11.67, 13.17
Lively	20.39, 14.94	27.28, 34.35	11.44, 10.01
Beautiful	17.30, 24.05	30.42, 23.82	13.10, 13.84
Wealthy	20.58, 25.94	29.85, 27.31	10.83, 12.21
Boring	22.19, 19.96	26.87, 29.83	14.00, 11.16
Depressing	23.94, 22.44	25.00, 27.48	11.10, 13.49

accordance with the common rules to obtain the texture information distribution of each image. Second, we use the trick in Subsection 4.2 to generate the ranking scores of all street images contained in the dataset using the AR-CNN3 model that is pretrained on six different perception attributes and combined with the TrueSkill algorithm. The ranking scores and texture distribution of the attributes of an image are calculated, and the results are shown in Table 4. This table shows that the safer the human feels, the higher is the proportion of edges. By contrast, the more insecure the human feels, the higher the proportion of flats. The more beautiful the human feels, the higher the proportion of flats, and vice versa. The more bored the human feels, the higher the proportion of edges. By contrast, the less bored the human feels, the higher the proportion of flats.

4.3.3 *Visual words*

The purpose of visual word analysis is to visualize the relationship between visual elements and the human perception of the environment. The dataset used in our work consists of street images from many cities on different continents and is similar to the dataset used by Zhou et al. [31]; thus, we use the Places-CNNs model of [31] to determine the scene attributes and categories of all images contained in the dataset, including fields. The visual vocabulary used in the scene category has 356 categories, and the number of scene attribute tags in SUNDASTAT exceeds 900. Our purpose is to understand the relationship between perceptual attributes and scene attributes from the visual vocabulary. Then, we use the trick in Subsection 4.2 to generate the ranking scores of all street images contained in the dataset using the AR-CNN3 model that is pretrained on six different perception attributes and combined with the TrueSkill algorithm. On this basis, we combine the prediction score statistics of the image to obtain the following results: The scene category distribution of the high-score images obtained by our prediction of safe attributes is mainly promenade, street, downtown, arcade, driveway, and residential neighborhood, and the scene attributes are concentrated in man-made and driving areas. Furthermore, the scene category distribution predicted from low-score images are mainly alleys, villages, and slums, and the scene attributes are concentrated in pavement areas. For the lively attribute, the scene category distribution of high-score images is mainly bazaars, bus stations, and crosswalks, and the scene attributes are concentrated in asphalt and man-made areas. For the beautiful attribute, the scene category distribution of high-score images is mainly parks, promenades, and campuses, and the scene attributes are trees and glass. The scene category distribution of high-score images in terms of the wealthy attribute is mainly office buildings and downtown, and the scene attributes are concentrated in shopping, open, and man-made areas. The scene category distribution of high-score images in terms of the boring attribute is farms, badlands, and field roads, and its scene attributes are open areas, no horizons, and railroads. The scene category distribution of high-score images in terms of the depressing attribute is construction sites and slums, and its scene attributes focus on no horizons and bricks.

5 Conclusion

In this study, we present a ranking network model based on an attention mechanism. The advantage of the model's attention mechanism is that the training of the original image is improved by the mixed attention feature representation of the color and texture attention mechanism. This advantage comes from the attention mechanism that is incorporated into the original image training through specific color- and texture-processed images during the training process. The influence of color and texture information on the ranking of the original image is also analyzed in the experimental part of this study. The experimental

results show that our method achieves state-of-the-art performance in comparison with the RSS-CNN algorithm.

Our technique can be generalized for computer vision tasks of studying perception or the visual attributes of images or scene categories. Our analysis of visual words enables the study of various research questions, such as How does semantic information affect residents' perceptions of urban appearance? How are different semantic objects of images perceived?

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant No. 62032020), in part by Hunan Science and Technology Planning Project (Grant No. 2019RS3019), in part by Hunan Provincial Natural Science Foundation of China for Distinguished Young Scholars (Grant No. 2018JJ1025), and in part by Guangzhou Research Project (Grant No. 201902010037).

References

1 Wilson J Q, Kelling G L. Broken windows. Atl Mon, 1982, 249: 29–38
2 Salesses P, Schechtner K, Hidalgo C A. The collaborative image of the city: mapping the inequality of urban perception. PLoS ONE, 2013, 8: 68400
3 Dubey A, Naik N, Parikh D, et al. Deep learning the city: quantifying urban perception at a global scale. In: Proceedings of European Conference on Computer Vision, 2016. 196–212
4 Naik N, Philipoom J, Raskar R, et al. Streetscore-predicting the perceived safety of one million streetscapes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014. 779–785
5 Ren J, Shen X H, Lin Z, et al. Personalized image aesthetics. In: Proceedings of IEEE International Conference on Computer Vision, 2017. 638–647
6 Dhar S, Ordonez V, Berg T L. High level describable attributes for predicting aesthetics and interestingness. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011. 1657–1664
7 Isola P, Xiao J X, Torralba A, et al. What makes an image memorable? In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011. 145–152
8 Quercia D, O'Hare N K, Cramer H. Aesthetic capital: what makes London look beautiful, quiet, and happy? In: Proceedings of ACM Conference on Computer Supported Cooperative Work and Social Computing, 2014. 945–955
9 Gibson J J. The ecological approach to visual perception. Science, 1979, 42: 98–99
10 Tamura H, Mori S, Yamawaki T. Textural features corresponding to visual perception. IEEE Trans Syst Man Cybern, 1978, 8: 460–473
11 Liu J L, Lughofer E, Zeng X Y. Aesthetic perception of visual textures: a holistic exploration using texture analysis, psychological experiment, and perception modeling. Front Comput Neurosci, 2015, 9: 134
12 Thompson M, Haber R N, Hershenson M. The psychology of visual perception. Leonardo, 1976, 9: 74
13 Trussell H J, Lin J, Shamey R. Effects of texture on colour perception. In: Proceedings of the 10th IVMSP Workshop: Perception and Visual Signal Analysis, 2011. 7–11
14 Chapelle O, Keerthi S S. Efficient algorithms for ranking with SVMs. Inf Retrieval, 2010, 13: 201–215
15 Ordonez V, Berg T L. Learning high-level judgments of urban perception. In: Proceedings of European Conference on Computer Vision, 2014. 494–510
16 Porzi L, Samuel R B, Lepri B, et al. Predicting and understanding urban perception with convolutional neural networks. In: Proceedings of ACM International Conference on Multimedia, 2015. 139–148
17 Radenović F, Tolias G, Chum O. Fine-tuning CNN image retrieval with no human annotation. IEEE Trans Pattern Anal Mach Intell, 2019, 41: 1655–1668
18 Comaniciu D, Meer P. Mean shift: a robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell, 2002, 24: 603–619
19 Smith A R. Color gamut transform pairs. SIGGRAPH Comput Graph, 1978, 12: 12–19
20 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. 886–893
21 Jain A K, Farrokhnia F. Unsupervised texture segmentation using Gabor filters. Pattern Recogn, 1991, 24: 1167–1186
22 Yin W, Schütze H, Xiang B, et al. ABCNN: attention-based convolutional neural network for modeling sentence pairs. In: Proceedings of the Transactions of the Association for Computational Linguistics, 2016. 259–272
23 Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 2204–2212
24 Fu J L, Zheng H L, Mei T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4438–4446
25 Chen Q, Hu Q M, Huang J X, et al. Enhancing recurrent neural networks with positional attention for question answering. In: Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017. 993–996
26 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of International Conference on Learning Representation, 2015
27 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 5998–6008
28 Chorowski J, Bahdanau D, Serdyuk D, et al. Attention-based models for speech recognition. In: Proceedings of Advances in Neural Information Processing Systems, 2015. 577–585
29 Wang F, Jiang M, Qian C, et al. Residual attention network for image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3156–3164
30 Herbrich R, Minka T, Graepel T. TrueSkill: a Bayesian skill rating system. In: Proceedings of Advances in Neural Information Processing Systems, 2007. 569–576
31 Zhou B L, Lapedriza A, Khosla A, et al. Places: a 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 1452–1464