International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-021-01401-w

Deep semantic-aware network for zero-shot visual urban perception

Chunyun Zhang[1] · Tianze Wu[2] · Yunfeng Zhang[1] · Baolin Zhao[3] · Tingwen Wang[1] · Chaoran Cui[1] · Yilong Yin[4]

Received: 1 December 2020 / Accepted: 22 July 2021

Abstract
Visual urban perception has recently attracted a lot of research attention owing to its importance in many fields. Traditional methods for visual urban perception mostly need to collect adequate training instances for newly-added perception attributes. In this paper, we consider a novel formulation, zero-shot learning, to free this cumbersome curation. Based on the idea of different images containing similar objects are more likely to possess the same perceptual attribute, we learn the semantic correlation space formed by objects semantic information and perceptual attributes. For newly-added attributes, we attempt to synthesize their prototypes by transferring similar object vector representations between the unseen attributes and the training (seen) perceptual attributes. For this purpose, we leverage a deep semantic-aware network for zero-shot visual urban perception model. It is a new two step zero-shot learning architecture, which includes supervised visual urban perception step for training attributes and zero-shot prediction step for unseen attributes. In the first step, we highlight the important role of semantic information and introduce it into supervised deep visual urban perception framework for training attributes. In the second step, we use the visualization techniques to obtain the correlations between semantic information and visual perception attributes from the well trained supervised model, and learn the prototype of unseen attributes and testing images to predict perception score on unseen attributes. The experimental results on a large-scale benchmark dataset validate the effectiveness of our method.

Keywords Visual urban perception · Semantic awareness · Zero-shot learning · Deep learning

1 Introduction

Visual urban perception [10, 30, 34, 40] aims to quantify the associations between the physical appearance of urban environment and the perceived feelings of its inhabitants (e.g., safety, wealth, and beauty). The famous *Broken Windows Theory* [54] pointed out that the visual signs of environmental chaos, including broken windows, littering, wandering, and graffiti, can significantly cause negative social consequences and increase the crime rates. Social scientists have also found the evidence about the impact of the visual quality of urban space on physical activities [18, 39], health [5], and education [33]. Hence, it becomes critically important to understand people's perceptions and evaluations of urban spaces.

Generally, visual urban perception is accomplished by a series of procedures such as interviewing urban residents and manually viewing photos. Undoubtedly, this is a tedious process that requires a lot of joint collaborations [36]. Fortunately, current large-scale geo-marked images obtained from online geographic communities have been regarded

✉ Chaoran Cui
crcui@sdufe.edu.cn

Chunyun Zhang
zhangchunyun1009@126.com

Tianze Wu
wtz@qlu.edu.cn

Yunfeng Zhang
qyhjs@163.com

Baolin Zhao
zhaobaolin-1@163.com

Tingwen Wang
521wtw@163.com

Yilong Yin
ylyin@sdu.edu.cn

[1] School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China

[2] School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

[3] Inspur Information Technology Co., Ltd., Jinan, China

[4] School of Software, Shandong University, Jinan, China

Published online: 17 August 2021

as a valuable resource for recent research on visual urban perception [10, 60, 63]. Also, the popularity of some crowdsouring platforms (e.g., Amazon MTurk) has made it convenient to collect the opinions of millions of users about urban areas. The above two factors can significantly promote the development of intelligent computing methods on predicting the high-level perceptual attributes of the urban environment. Early visual urban perception was initially considered as a regression problem [35, 45], which assigns a score related to specific perception attributes for images in different locations of the city. For human observers, it is quite difficult for people to give an absolute evaluation about images in daily life, while relative judgements (e.g. "*does this image look safer than the other one?*") are more natural [6, 30, 40]. Therefore, recent research [21, 60] is dedicated to directly predicting the relative ranking of two scene images with respect to different perceptual attributes. However, these methods can not answer the question "*what makes the image look safe?*", which is important for the newly-added perception attribute prediction without training examples.

Intuitively, visual urban perception is accompanied by the recognition of semantic content of images [20]. Before people describing the image of urban space as lively, boring or beautiful, they need to understand what they have seen. To better illustrate this point, we obtain attribute activation maps to localize the discriminative image regions for visual urban perception [59, 62] by using a standard Convolutional Neural Network (CNN) (the details of this process will be discussed in Sect. 4.1.1). Figure 1 shows the attribute activation maps of different example images with respect to different perceptual attributes. As can be seen from the Fig. 1, the discriminative regions are usually the parts that contain some semantic objects in the image. For example, the activation parts regarding the attribute "*Safety*" is a fence, and the magnificent building plays a critical role in the perception of the attribute "*Wealthy*". Such results confirm that visual urban perception needs to be accompanied by the semantic understanding of urban space. According to this knowledge, if we introduce semantic understanding features into our proposed visual urban perception, then we can answer the question "*what makes the image look safe?*".

Moreover, with the growing advancement of multimedia and computer vision technologies, many works [35, 40, 42, 60] achieve notable performance on automatic special urban perceptual attribute prediction. However, for newly-added

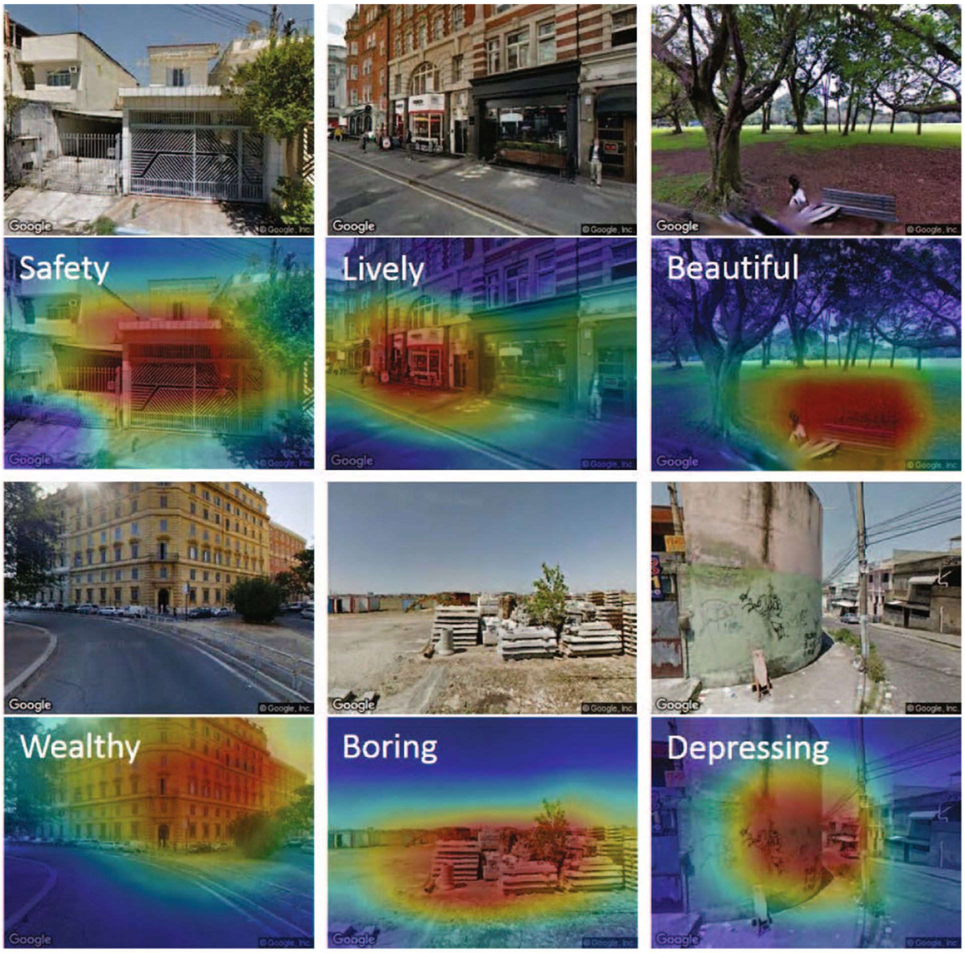

Fig. 1 The activation maps for images on six perceptual attributes

attribute prediction, there is still a non-negligible limitation. Obtaining adequate training instances for every new attribute is an increasingly impractical solution. Therefore, people prefer an automatic completion solution, or even a more radical method that recognizes unseen attributes without seeing any training instances.

Recently, zero-shot learning has received a growing amount of attention [26, 37, 53, 57, 58]. It aims to recognize objects or facts of new classes or attributes (unseen classes or attributes) with no examples being seen during the training stage. A key to zero-shot learning is to construct a shared semantic space between seen classes and unseen classes. Generally speaking, based on how a semantic space is constructed, zero-shot methods are categorized into engineered semantic space based method and learned semantic space based method [53]. For engineered semantic space based method, each dimension of the semantic space is designed by humans [11, 23]. Hence, it has poor domain adaptability. In learned semantic space based method, the dimensions in the spaces are not designed by humans. Prototypes of each class are obtained through the embedding of class labels [32, 46, 56], the text descriptions for each class [12, 19, 27, 47] or images belonging to each class [52, 57, 58]. The learned semantic space based method can generate semantic spaces containing information that can easily overlooked by humans.

Analogous to the learned semantic space based paradigms, we try to construct a shared semantic space by learning. According to the conclusion of the semantic activation map, different images containing similar objects are more likely to possess the same perceptual attribute. From this standpoint, suppose if we incorporate semantic information and perception information to form a new fusion semantic space, and in this new space we represent an image sample by a vector indicating the probability of the objects, then we can compare the object vectors represented in the same semantic correlation space. Hence, we define the zero-shot learning for visual urban perception prediction problem as a two-step problem: train a supervised deep semantic-aware perception network for training attribute step and adopt the zero-shot learning method to predict unseen perception attributes step. In the first supervised training step, we propose a double-column CNN architecture to predict the perception of images on specific perceptual attributes through paired image comparison, and generate attribute activation maps to vividly display the relationship between object categories and perceptual attributes. After that, we add the semantic information of the image into each column of the original neural network to train a semantic-aware network for visual urban perception. In the second step, we use the visualization of network to obtain the semantic correlation matrix between objects and visual perceptual attributes from the well trained supervised model in the first step. Based on the semantic correlation, we can synthesize unseen attribute prototypes by calculating similar object vector representations between the unseen attributes and the training perceptual attributes. At last, we represent the test images and the unseen attributes over the space, and utilize simple similarity algorithm to predict visual urban perception for the test images on the unseen attributes. This means that we incorporate the semantic information with general perception features to facilities the final perception result in supervised visual urban perception prediction network, and based on the well trained supervised prediction network to learn the connection of seen classes and unseen classes in a new fusion semantic space with object semantic information as bridge.

Generally speaking, compared with traditional zero-shot learning methods,our proposed two-step zero-shot learning paradigm has two added benefits. Firstly, compare with traditional methods, the two-step zero-shot learning method is easier to transition from the traditional supervised VUP method. Since most VUP methods are trying to improve their performance in traditional paradigm, we can conduct our zero-shot step on these well trained state-of-the-art methods to predict unseen perception attributes. Secondly, we leverage a visualization of network to acquire the intrinsic relationships between semantic information and perceptual attributes from the pre-trained supervised VUP prediction model. It shows that our proposed two-step method is more interpretable compared with traditional zero-shot learning methods. The main contributions of our work can be summarized as follows:

- We proposed a novel two-step zero-shot learning method for visual urban perception. In the first step, it trains a supervised semantic-aware perception prediction model for training attributes . In the second step, it learns a semantic shared space of seen and unseen attributes from the well trained model. Based on the shared semantic space, we can predict perception score of newly-added attributes.
- We illustrate the importance of semantic information of images in visual urban perception and we combine semantic information with the generic features of images to form our supervised semantic-aware perception prediction framework. This means that our proposed model can answer questions such as “what makes the image look safe?”.
- We leverage the visualization of network to acquire the intrinsic relationships between semantic information and perceptual attributes from the pre-trained neural network, which further improves the interpretability of our two-step zero-shot learning architecture.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 gives the problem definition. Section 4 introduces the proposed two-step zero-shot learning method for visual urban perception in detail. The experiments conducted on a large-scale benchmark dataset are presented in Sect. 5. The conclusion and future work are pointed out in Sect. 6.

2 Related work

In this section, we review key works from two research fields: supervised urban visual perception and zero-shot learning.

2.1 Supervised visual urban perception

Visual urban perception has received increasing attention due to its far-reaching role in numerous fields. Therefore, there has been a lot of research work on urban visual perception in recent years [8, 10, 15, 16, 60, 61, 63]. Generally, the prediction methods for visual urban perception can be categorized into two classes: feature based method and neural network based method.

In feature-based approaches, different sets of features are extracted and fed into a chosen regressor(e.g. Linear Regression). Naik et al. [35] tried to input generic image features into support vector regression to predict the safety perception of street-level images. Ordonez and Berg [38] achieved comparable outcomes by using Fisher vectors and DeCAF features to predict safety and wealthy perception. However, feature-based approaches are mostly based on traditional machine learning methods. They largely depend on human feature engineering and are difficult to migrate to other scenarios.

With the prosperity of deep learning recently, neural networks have been widely used to automatically learn visual perception features, and achieved the state-of-the-art performance. In neural network based method, Convolution Neural Network (CNN) is widely used model in visual urban perception [17, 49, 60]. Porzi et al. [40] utilized CNN model to identify the mid-level visual elements and obtained an attractive result on the perception of safety. Dubey et al. [10] designed a dual-column CNN with tied weights for capturing image features and predicting the image's score on specific perceptual attributes via pairwise comparison. Liu et al. [30] introduced a deep CNN and an EM algorithm to quantify the perceptual degree of the urban physical environment using unlabeled crowdsourced street view data.

Additionally, Law et al. [25] dedicated to assessing house prices through the perceptual features extracted by CNN. All the above work is mainly based on the CNN network to obtain perceptual features. Our proposed supervised visual urban perception model is based on CNN framework, but different from the previous researches, we utilize a double-column CNN architecture. For the double column CNN architecture, named DCPN in our paper, it consists of two identical columns based on ResNet-50 [13] which replaces the last layer containing 1000 neurons with neuron. Each column takes one input image and outputs one real value which reflects the degree of its perceptual attribute. Then the two output values are input to the last layer for comparison. Additionally, the DCPN is different from dual-column CNN proposed by Dubey et al. [10]. Beside the final ranking loss, the dual-column CNN includes a fusion sub-network and use softmax loss to train the two basic model based on AlexNet, while the DCPN only use the ranking loss with the Resnet-50 as the basic model pre-trained on Image-Net.

A series of studies has pointed out that the semantic information of images has significant influence on urban visual perception [3, 14]. For example, He et al. [14] demonstrated that compared to some ordinary ones, buildings with dominant shapes and bright colors, historical sites and intrusive signs are more attractive to humans. Can et al. [3] also proved that different objects contained in an image may represent distinct atmospheres. However, they didn't conduct further research on the effect of semantic information on image perception. In our work, we prove the importance of semantic information in urban perception by generating the attribute activation map. Further more, we leverage the object semantic information to assist perception prediction and experimental result proved that it indeed improves the accuracy of the visual perception task.

In addition, to further illustrate the correlation of semantic features and the prediction target, Mahendran and Vedaldi [31] and Dosovitskiy and Brox [9] analyzed the visual encoding of CNNs by inverting deep features at different layers, which only shows the information retained in the deep features without highlighting the relative importance of this information. Zhou et al. [62] identified the importance of the image regions by projecting back the weights of the output layer on to the convolutional feature maps. Inspired by their work, we introduce the visualization techniques to highlight exactly which regions of an image are important for discrimination on specific perceptual attribute.

2.2 Zero-shot learning

Supervised classification methods have achieved significant success in research and have been applied in many areas. However, these methods need sufficient labeled training instances and can not deal with previously unseen classes. To solve this problem, zero-shot learning methods are proposed. The aim of zero-shot learning is to classify instances belonging to the classes that have no labeled instances. Since its inception [23, 24], zero-shot learning has become a fast-developing research direction in machine learning, and is widely used in computer vision and natural language processing.

A key component of zero-shot learning is how to construct the shared semantic space for seen classes and unseen classes. Various semantic spaces have been exploited by existing works. According to how a semantic space is constructed, zero-shot learning methods are categorized in two classes: engineered semantic space based method and learned semantic space based method [28, 53]. In the former method, the semantic space is designed by humans with a set of attributes [11, 29], lexical items [1] or keywords extracted from the text descriptions of each class [41]. However, engineered semantic space based method has poor domain adaptability since it largely depends on humans to design the semantic space. In the latter method, prototypes of each class are obtained from the output of some machine-learning models. In these prototypes, each dimension does not have an explicit semantic meaning. Instead, the semantic information is contained in the whole prototype. The models used to extract the prototypes can be pre-trained in other problems or trained specifically for the zero-shot learning problem [53]. Benefit from the development of unsupervised neural language model [2, 7, 51], most learned semantic space based methods construct semantic space through the embedding of class labels [12, 46, 50, 56], or the text descriptions for each class [12, 19, 27, 47]. Besides, there are many learned semantic spaces learned from images belonging to each class [52, 57, 58]. The learned semantic space based method can largely reduce manual participation and generate semantic spaces containing information that are easily overlooked by humans.

Analogous to the learned semantic space based paradigms, in this paper, we propose a new zero-shot learning method for visual urban perception. Different from traditional methods constructing semantic space only through semantic information, we learn the shared semantic space between the seen attributes and unseen attributes by incorporating semantic information with task-specific information (general perception features). The proposed zero-shot learning method includes two steps: supervised visual urban perception prediction model training and zero-shot learning for newly-added attributes. In the supervised visual urban perception model training step, we can obtain a semantic correlation matrix (SCM) from the well trained model by incorporating semantic information with general perception information. In the zero-shot learning step, based on the learned SCM, we can synthesize unseen attribute prototypes and testing instance prototypes and further conduct zero-shot learning for newly-added attributes prediction. Note that, since we are more concerned about how to construct a semantic correlation matrix, we only use WordNet [43] to learn the attribute label embedding instead of using embedding or the text descriptions of attribute labels from unsupervised methods [19, 47].

Note that this paper is an extension of our conference version in [59]. In this paper, we provide a more comprehensive literature review on supervised visual urban perception and zero-shot learning methods. More importantly, due to the difficulty of collecting labeled images for visual urban perception, we propose a two-step zero-shot learning method to predict the degree of perception for images on unseen attributes by learning a shared semantic space from a trained deep semantic-aware visual urban perception model. We also conduct experiments to validate the effectiveness of the proposed method in comparison with other zero-shot learning schemes.

3 Problem definition

In this paper, we formulate visual urban perception as a two-step zero-shot learning problem: for seen attributes training, we formulate it as a supervised visual urban perception problem, which tries to supervised train a deep semantic-aware visual urban perception model; for unseen attributes, we can learn shared semantic space between seen and unseen attributes from the well trained model and further predict perception scores of unseen attributes. Hence, we first present the problem definition of supervised visual urban perception for seen attributes, and then give some notations of zero-shot learning for newly-added attribute perception prediction.

3.1 Supervised visual urban perception

In the supervised visual urban perception, we focus on the relative order relationships between different images, and regard the urban perception prediction problem as a ranking task. Formally, given the pairwise comparison results among images on specific perceptual attributes, each pairwise comparison can be seen as a triple $(\mathbf{x}_i, \mathbf{x}_j, y)$, where $\mathbf{x}_i, \mathbf{x}_j \in \boldsymbol{X}$ are images and $y \in \{+1, -1\}$ represents the label. Taking "*Boring*" attribute as an example, $y = +1$ denotes that image $\mathbf{x}_i$ is more boring than image $\mathbf{x}_j$, and $y = -1$ means the reverse. Therefore, a given dataset with n pairwise comparisons is defined as $\boldsymbol{D} = \left\{(\mathbf{x}_i^k, \mathbf{x}_j^k, y^k)\right\}_{k=1}^{n}$. The goal of our task is to learn a mapping function $\boldsymbol{F}_z : \boldsymbol{X} \to \mathbb{R}$ that predicts a value $\boldsymbol{F}_z(\mathbf{x}_i)$ for each image $\mathbf{x}_i$ perceived by people on the specific perceptual attribute z. For example, $\boldsymbol{F}_{Boring}(\mathbf{x}_i)$ means the degree of boredom of $\mathbf{x}_i$ perceived by people. The desired mapping function $\boldsymbol{F}_z$ is collected by minimizing the following hinge loss function:

$$\boldsymbol{L} = \sum_{(\mathbf{x}_i, \mathbf{x}_j, y) \in \boldsymbol{D}} \max(0, y \cdot (\boldsymbol{F}_z(\mathbf{x}_j) - \boldsymbol{F}_z(\mathbf{x}_i)) + 1). \tag{1}$$

3.2 Zero-shot learning for visual urban perception

Assuming $\mathbf{Z}_{seen}$ as the attribute set of training instances and $\mathbf{Z}_{unseen}$ as the attribute set of testing instances, where $\mathbf{Z}_{seen} \cap \mathbf{Z}_{unseen} = \emptyset$. Zero-shot learning aims to transfer knowledge from seen (source) attributes to disjoint set of unseen (target) attributes and improve the predictive accuracy of unseen attributes by sharing the semantic space with seen attributes.

Initially, different images containing similar objects are more likely to possess the same perceptual attribute, which is verified in Sect. 4.1.1. That is to say, the semantic correlation space formed by objects semantic information and perceptual attributes can accurately measure the semantic distance of images, which is the key basis that we rely on for zero-shot learning. From this standpoint, if an image sample can be represented by a vector indicating the probability of the objects, we can compare the object vectors represented in the same semantic correlation space. Hence, based on the pre-trained model, we can obtain the object-aware Semantic Correlation Matrix (SCM) $\prod$ by using the visualization techniques of the network [48] to obtain the correlations between objects and perceptual attributes. Here, we assume that we have learned enough object semantics at the semantic level. The semantic correlation matrix $\prod$ provides the prototypes of all the training perceptual attributes, but for an unseen attribute without training samples, its prototypes cannot be obtained directly by learning. However, with the obtained SCM and a simple similarity metric between seen and unseen attributes, we can synthesize unseen attribute prototypes by transferring similar object vector representations between the unseen attributes and the training perceptual attributes.

Hence, given the synthesized unseen attribute prototype $\prod_{\hat{z}}$ and the representation of the testing image $g(I)$, the perceived degree of test image I on the unseen attribute $\hat{z}$ can be calculated by the following formula:

$$\boldsymbol{F}_{\hat{z}}(I) = cos\left(g(I), \prod\nolimits_{\hat{z}}\right) \tag{2}$$

where $\hat{z} \in \mathbf{Z}_{unseen}$, $cos(\cdot)$ represents the cosine similarity.

Since we assume that we have learned enough object semantics at the semantic level, we will only further discuss how to learn representation of the unseen attribute prototype $\prod_{\hat{z}}$ and the testing image $g(I)$ in Sect. 4.2.2.

4 Framework

As mentioned above, we formulate the visual urban perception problem as a two-step zero-shot learning problem: supervised visual urban perception step for training attribute and zero-shot prediction step for newly-added attributes. Hence, in this part, we first introduce activation map technique to highlight the most discriminative regions of image on specific perceptual attributes. Based on this knowledge, we propose a deep semantic-aware perception network that relies on semantic recognition to assist supervised urban perception prediction. Then, we show how to use the well trained supervised model to obtain the semantic correlation matrix of the perceptual attributes. Finally, based on the semantic correlation matrix, we obtain unseen attribute prototype and testing image prototype to do zero-shot prediction on newly-added attributes.

4.1 Supervised visual urban perception training

4.1.1 Attribute activation map

In this section, we try to use the attribute activation map to verify that semantic information is helpful for visual urban perception prediction and answer the question "*what makes the image look safe?*".

Based on this motivation, we first propose a Double-Column Perceptual Network (DCPN) with a pair of images as input to solve the supervised visual urban perception problem. The architecture of the DCPN is illustrated in Fig. 2. The DCPN consists of two columns of Perception Stream (P-Stream) based on ResNet-50 [13]. At the last layer of P-Stream, we replace 1000 neurons with 1 neuron outputting the perceived degree of the input image on a specific attribute.

In order to explore which parts of the image affect the perception result of urban, we then apply activation mapping technique [62] to highlight the most distinctive image areas on specific attributes. For the well trained DCPN, we choose one column to generate activation maps. The DCPN mainly consists of several convolution layers, a global average pooling layer, and a fully connected layer successively. Given an input image, after a block of convolution layers, the network can generate feature maps, where K is the number of filters in the last convolution. On the top of these convolution layers, global average pooling layer takes the average of each feature map and generates a K-dimension vector $W = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k)$. Then, the resulting vector connects to the fully connected layer. In our network, the fully connected layer has just one neuron corresponding to the degree of one perceptual attribute. Notably, each value of K-dimension vector is generated by the global average pooling layer that outputs the spatial average of each feature map. Therefore, the set of $W = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k)$ reflect the importance of each feature map of K feature maps too. In fact, the activation map for a specific attribute is the weighted sum of K feature maps outputted by the last convolution layer, while the weight of each feature map indicates its importance for that attribute.

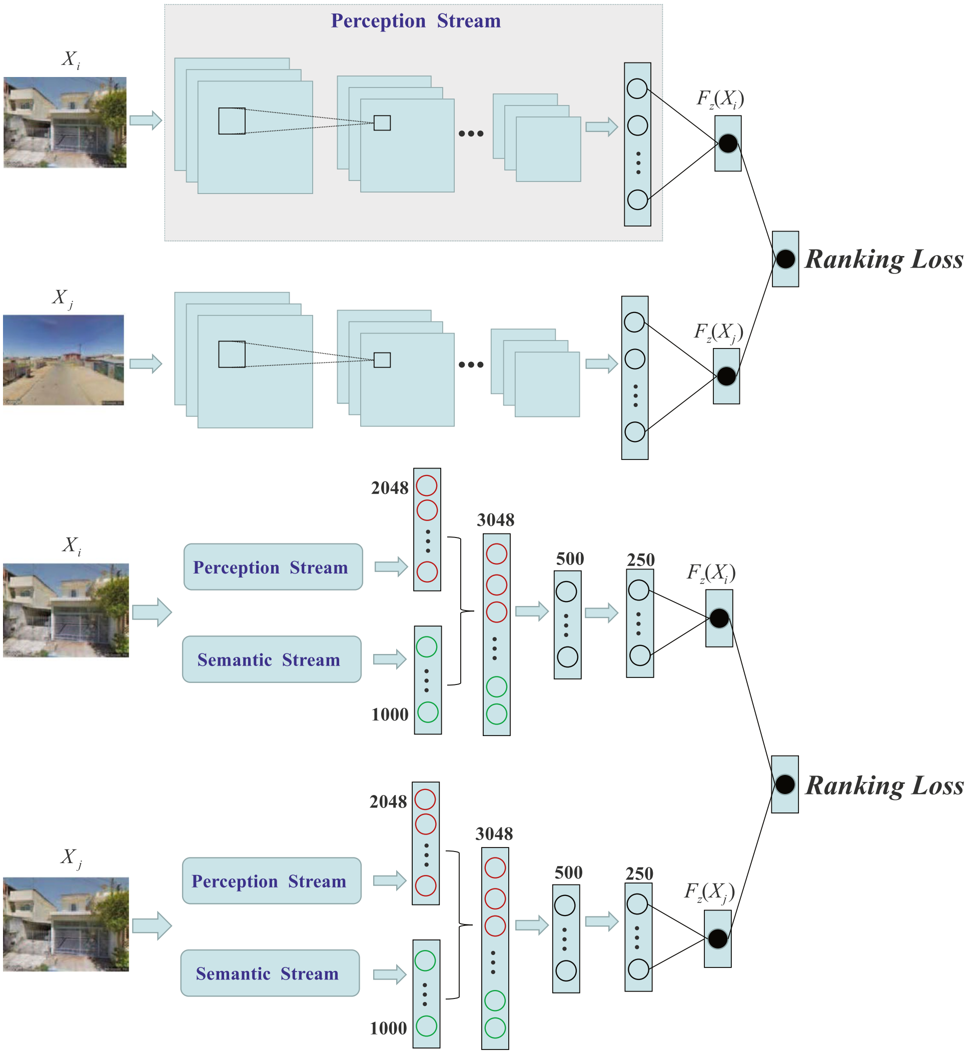

Fig. 2 The network architecture comparison between DCPN (above) and DSANZS (below)

Formally, as shown in Fig. 3, given an image $\mathbf{x}_i$, $\boldsymbol{A}_k(\mathbf{x}_i)$ denotes the k -th feature map for image $\mathbf{x}_i$. The activation map $\boldsymbol{M}(\mathbf{x}_i)$ of the image $\mathbf{x}_i$ is calculated by:

$$\boldsymbol{M}(\mathbf{x}_i) = \sum_{k=1}^{K} \mathbf{w}_k \boldsymbol{A}_k(\mathbf{x}_i) \tag{3}$$

where $\mathbf{w}_k$ is equivalent to the value of the k-th unit of the vector generated by the global average pooling layer (to simplify the notation here, we ignore the bias term).

Figure 1 displays several activation maps of images on six different perceptual attributes. It can be seen from the picture, the region of image that best reflects "*Wealthy*" attribute is magnificent building, while the region of image that best reflects "*Depressing*" attribute is graffiti. In other words, semantic information such as magnificent building and graffiti are highly correlated with both "*Wealthy*" and "*Depressing*" separately, which is also consistent with people's cognitive behavior. From the above observations, it can be concluded that the semantic information in the image is indeed beneficial to the task of urban visual perception. Further, through the activation maps for images on perceptual attributes, we can answer the question"*what makes the image look safe or wealthy?*".

4.1.2 Deep semantic-aware perception network

Motivated by the finding in Sect. 4.1.1, we simultaneously perform the tasks of semantic recognition and visual urban perception to boost the performance of visual urban perception. We introduce a Semantic Stream (S-Stream) to the DCPN model to extract the semantic information of images [55]. We name it as Deep Semantic-Aware Perception Network (DSAPN), whose network structure is shown in Fig. 2.

As shown in Fig. 2, given an image, we use P-Stream and S-Stream to extract 2048-dimension general features and 1000-dimension semantic features separately, and then concatenate them to obtain 3048-dimension fusion features as the input to a fusion network with three fully connected

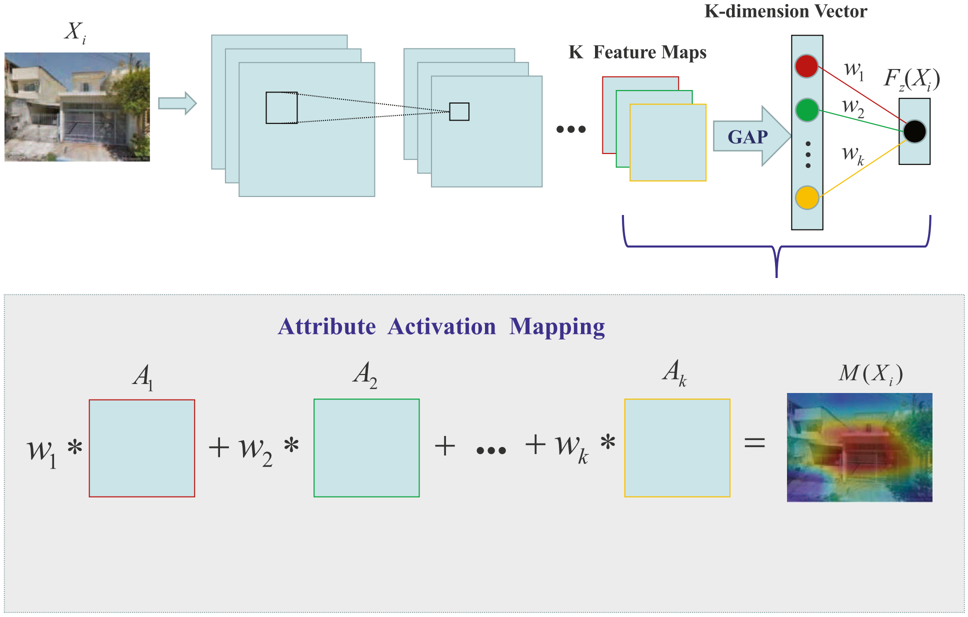

Fig. 3 Attribute activation mapping generation process

layers. At last, a real value is outputted reflecting the perceived degree of the input image on a specific attribute. Note that the semantic stream, perception stream and fusion network is defined as following:

- **Semantic Stream** aims to extract the semantic features of images. Specifically, this stream is based on a ResNet-50 model pre-trained on the ImageNet dataset, which consists of 1000 categories dedicated to classification [44]. The output of the last layer of the network ($FC1000$) is the input of the fusion network, in a nutshell, for the image $\mathbf{x}_i$, this stream outputs $\mathbf{x}_i^S \in \mathbb{R}^{1000}$.
- **Perception Stream** aims to extract generic features of images, which directly related to visual perception prediction. Similarly, this stream uses the pre-trained ResNet-50 as its basic model, and takes the output of the second-to-last layer as the input of the fusion network. It should be noted that Block-4 and Block-5 in ResNet-50 are fine-tuned during training. For the image $\mathbf{x}_i$, this stream outputs $\mathbf{x}_i^P \in \mathbb{R}^{2048}$.
- **Fusion Network** aims to fuse features generated by S-Stream and P-Stream and predicts the degree of perception for each image by utilizing a three-layered fully connected subnetwork.

Formally, given an image $\mathbf{x}_i$, the semantic features $\mathbf{x}_i^S$ and perceptual features $\mathbf{x}_i^P$ are extracted by S-Stream and P-Stream respectively, which are regarded as the input of the first hidden layer with 500 neurons in the fusion network. Next, the second hidden layer with 250 neurons performs dimensionality reduction and fusion processing on the outputs of the first hidden layer. Finally, the fully connected layer with just one neuron outputs a score reflecting the degree of the image on a specific perceptual attribute as $f_z(\mathbf{x}_i^S, \mathbf{x}_i^P)$. In other words, the output $f_z(\mathbf{x}_i^S, \mathbf{x}_i^P)$ of the fusion network for image $\mathbf{x}_i$ represents the degree of $\mathbf{x}_i$ perceived by people for corresponding perceptual attribute z. Thus, the hinge loss function (see Eq. (1)) can be redefined as:

$$\begin{aligned} \boldsymbol{L} &= \sum_{(\mathbf{x}_i, \mathbf{x}_j, y) \in \boldsymbol{D}} \max(0, y \cdot (\boldsymbol{F}_z(\mathbf{x}_j) - \boldsymbol{F}_z(\mathbf{x}_i)) + 1) \\ &= \sum_{(\mathbf{x}_i, \mathbf{x}_j, y) \in \boldsymbol{D}} \max(0, y \cdot (f_z(\mathbf{x}_j^S, \mathbf{x}_j^P) - f_z(\mathbf{x}_i^S, \mathbf{x}_i^P)) + 1). \end{aligned} \tag{4}$$

In our study, the S-Stream exploits a ResNet-50 network pre-trained on the ImageNet 2012 dataset consisting of 1000 object classes [44]. Therefore, the 1000-dimension vector extracted by S-Stream represents the probabilities of an input image belonging to different object categories. From this perspective, this 1000-dimension vector can be considered as semantic features of images. Different from S-Stream, P-Stream exploits a ResNet-50 network pre-trained on the target training dataset. The 2048-dimension feature vector extracted by P-Stream thus indicates the general features of images directly related to the task of visual urban perception.

4.2 Zero-shot learning for newly-added attribute perception prediction

Zero-shot learning aims to transfer knowledge from seen (source) attributes to disjoint set of unseen (target) attributes and improve the predictive accuracy of unseen attributes by sharing the semantic space with seen attributes. Based on the conclusion in Sect. 4.1.1, the semantic information in the image is indeed beneficial to the task of urban visual perception, we suppose that different images containing similar objects are more likely to possess the same perceptual attribute. From this point of view, in this section, we firstly learn the semantic correlation space formed by objects semantic information and perceptual attributes from the well trained DSAPN model, and then based on the semantic correlation space, we learn representation of the unseen attribute prototype and the testing image for zero-shot learning.

4.2.1 Semantic correlation analysis

Based on the pre-trained DSAPN network, we utilize the visualization technique of the network to explore the correlations between objects and perceptual attributes [48]. Specifically, when the output score of the network is the largest, the correlation between the predicted perceptual attributes and the input semantic information is the highest, and we thereby obtain the object category that best represents the perceptual attribute. Such correlation representation identifies the most discriminative objects for the specified urban perception.

As discussed in Sect. 4.1.2, $f_z(\mathbf{x}_i^S, \mathbf{x}_i^P)$ represents the perceptual score of image $\mathbf{x}_i$ on perceptual attribute z. We combine $\mathbf{x}_i^S$ and $\mathbf{x}_i^P$ as $\bar{\mathbf{x}}_i$, so $f_z(\bar{\mathbf{x}}_i) = f_z(\mathbf{x}_i^S, \mathbf{x}_i^P)$ also represents the perceptual score of image $\mathbf{x}_i$ on perceptual attribute z. We obtain the L_2-regularized feature representation by maximizing $f_z(\bar{\mathbf{x}}_i)$ as follows:

$$\hat{\mathbf{x}}_z^S = \arg\max_{\mathbf{x}_i^S} f_z(\bar{\mathbf{x}}_i) - \lambda \left\| \mathbf{x}_i^S \right\|_2 , \tag{5}$$

where λ is the regularization parameter, and the semantic representation $\mathbf{x}_i^S$ is collected by back-propagation with randomly initialized $\bar{\mathbf{x}}_i$. Equation (5) means that the $\mathbf{x}_i^S$ who can make $f_z(\bar{\mathbf{x}}_i)$ (the perceptual score of image $\mathbf{x}_i$ on perceptual attribute z) the largest, is the best semantic representation $\hat{x}_i^S$. The L_2 regularization in Eq. (5) is used to constraint the size of the $\mathbf{x}_i^S$ and to prevent overfitting. Since the Eq. (5) is to maximize the objective function, we use the difference between the maximum perceptual score and L_2 regularization to represent the optimal semantic representation .

Correspondingly, we fix the network weights and use the gradient ascent method to obtain object representations associated with perceptual attributes. Finally, the object-aware Semantic Correlation Matrix (SCM) is obtained:

$$\prod = \left[\hat{\mathbf{x}}_z^{S^T} \right]_z . \tag{6}$$

4.2.2 Zero-shot prediction via SCM correlations

Based on the obtained SCM matrix $\prod$, we will discuss how to obtain the unseen attribute prototype and testing instance prototype.

- **Unseen Attribute Prototype**: We can synthesize unseen attribute prototype using the SCM matrix:

$$\prod_{\hat{z}} = \sum_{z=1}^{|\mathbf{Z}_{T_r}|} sim(\hat{z}, z) \cdot \prod_z \tag{7}$$

  where $z \in \mathbf{Z}_{T_r}$ and $\hat{z} \in \mathbf{Z}_{T_e}$, $sim(\hat{z}, z)$ is the semantic similarity between the unseen attribute and the training attribute. $\prod_z$ is the column of $\prod$ for perceptual attribute z. There are many similarity functions we can choose, here we use WordNet [43] to define the similarity function $sim(\hat{z}, z)$ between the training attributes and the unseen attributes.
- **Testing-Instance Prototype**: We define a test image as a testing-instance prototype with ConSE [37]. For each of the training attributes, we convert the scores $\boldsymbol{F}(I)$ of an image I to a probability distribution $p(\cdot)$. Hence the perceived score for image I_i on a perceptual attribute $z \in \mathbf{Z}_{T_r}$ can be represented as $p(z|\bar{\mathbf{x}}_i)$, and the probability sum of all training attributes is $\sum_{z=1}^{|\mathbf{Z}_{T_r}|} p(z|\bar{\mathbf{x}}_i) = 1$.

  We use z_t to denote the perceptual attribute with the tth highest probability for image I according to $p(\cdot)$, and $p(z_t|\bar{\mathbf{x}})$ is the perceptual probability for image I on perceptual attribute z_t. Thus, given the top T perceptual attributes, the testing-instance prototype of the test image I can be obtained by SCM matrix:

$$g(I) = \frac{1}{\Delta} \sum_{t=1}^{T} p(z_t|\bar{\mathbf{x}}) \cdot \prod_{z_t} \tag{8}$$

  where $\Delta = \sum_{t=1}^{T} p(z_t|\bar{\mathbf{x}})$ is a normalization factor, and $\prod_{z_t}$ is the column of $\prod$ corresponding to perceptual attribute z_t.

After obtaining the representation of the target unseen attribute and testing image, we can predict the perception score of the testing image on the newly-added target attribute by using Eq. (2).

5 Experiments

In this section, we present the experimental details in our evaluation. In particular, we compare our proposed method with some baselines on the challenging large-scale Place Pulse 2.0 (PP 2.0) dataset. Note that since we proposed a two-step zero-shot learning method for visual urban perception, we compared our proposed method in each step with corresponding state-of-the-art methods. The experimental results show the advantages of semantic information and the superiority of our method for visual urban perception.

5.1 Experimental setup

5.1.1 Datasets

The Place Pulse (PP) 2.0 dataset [10] is one of the most large-scale and challenging dataset for street-level image perception. It provides the pairwise comparisons on the perceptual attributes of different images and is a benchmark dataset for visual urban perception. The PP 2.0 consists of 110,988 street-level images from 56 major cities of 28 countries, and there are 1,169,078 pairwise comparisons with respect to 6 perceptual attributes. In detail, there are 370,134 pairwise comparisons on "*Safety*", 268,494 pairwise comparisons on "*Lively*", 166,823 pairwise comparisons on "*Beautiful*", 137,688 pairwise comparisons on "*Wealthy*", 114,755 pairwise comparisons on "*Depressing*", and 111,184 pairwise comparisons on "*Boring*". Figure 4 presents example images from the Place Pulse 2.0 dataset. According to [10], the ground-truth perceptual attributes of images are collected from online volunteers. Specifically, the volunteers were asked to watch a randomly-chosen pair of images slide by slide, and answer which one is better in terms of one of the six perceptual attributes. In the dataset, images were only labeled with one single category. Especially, we selected 104,529 images and 1,049,415 image pairs from PP 2.0 for pairwise comparison to evaluate the accuracy of different methods on six perceptual attributes.

5.1.2 Implementation details

For supervised visual urban perception, our implementation is based on the Keras deep learning framework and uses the mini-batch SGD algorithm during network training [4].

Specifically, both DCPN and DSAPN perception streams use ResNet-50 as the basic model and use its pre-trained weights on the Image-Net dataset. Subsequently, we fine-tune the Block-4 and Block-5 of the two model with a batch size of 32. It will result in a large amount of calculation and a long training time if we perform fine-tune on the whole network model. Furthermore, the bottom layers of neural networks tend to extract only local features which are irrelevant to high-level perceived tasks. In contrast, the top layers are able to extract global features that are highly correlated with the high-level perceived tasks. Therefore, we only fine-tune Block-4 and Block-5 during the training process. The learning rate and momentum are initialized as 10^{-3} and 0.9, respectively. When the validation error stopped improving with current learning rate, we reduced it by a factor of 10,

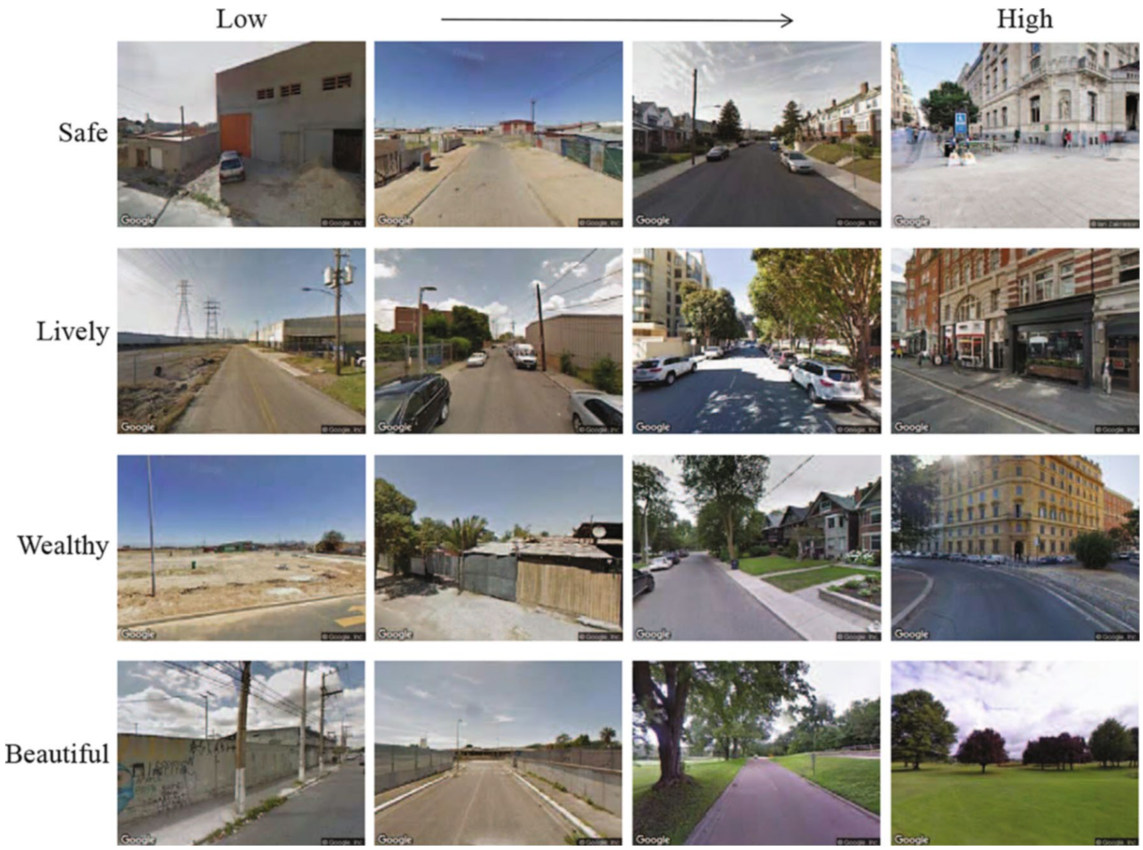

Fig. 4 Example images from the Place Pulse 2.0 dataset, which are ranked based on the predicted scores of the proposed DSAPN

repeating this process a maximum of four times (following [22]). We stipulate that the learning rate is reduced to 0.1 times of the original when the loss on validation set no longer drops in 10 epochs. The networks were trained to 100,000–150,000 iterations, stopping when the validation error stopped improving even after decreasing the learning rate. All biases are initialized with zero.

5.2 Supervised visual urban perception

We propose two network structures, DCPN and DSAPN, and compare DSAPN with a series of latest supervised methods [10] for visual urban perception to verify their effectiveness of our model. Particularly, using the pre-trained ResNet-50 on the ImageNet dataset and fine-tuned ResNet-50 on PP 2.0 as two basic models. Both PRN and DSAPN contain two identical P-Stream with the same parameters.

- **DCPN (ResNet)** takes an image pair as input and is composed of two ResNet-50 networks with pre-trained weights. We change 1000 neurons to one neuron in the last layer of the fully connected layer and all the weights in convolution layers are fixed. The network tries to find a mapping function to predict a perceived score for each image and combines two scores with hinge loss;
- **DCPN (fine-tuned)** has the same structure as DCPN (ResNet). Differently, we fine-tune the Block-4 and Block-5 of the ResNet-50;
- **DSAPN (ResNet)** combines generic features and semantic information for an image to predict a score for a specific perceptual attribute. The weights of the network in perception stream are fixed as the pre-trained weights;
- **SS-CNN** [10] takes a pair of images as input. One view of the network is that it consists of two parts: two disjoint identical sets of layers with tied weights for feature extraction, and a fusion sub-network with softmax loss used to train the network following the extractor layers;
- **RSS-CNN** [10] is the improvement of SS-CNN. It modifies the SS-CNN by adding a ranking sub-network, consisting of fully connected layer with tied weights. It combines softmax loss and ranking loss to train the network together.

In our experiments, the dataset is divided into training set, validation set and test set according to 65-5-30. Experimental results are illustrated in Table 1, 2 and 3. Note that the numbers marked in bold are the best experimental results of all methods. Table 1 details the accuracy results of different methods on PP 2.0. We take the average of five (empirical value is 5 or 10) experimental results as the final result. Since the code of the work [10] is not publicly available, we just implemented the basic versions of SS-CNN and RSS-CNN, without the post-processing for further ranking the scores with TrueSkill and RankSVM. The average prediction accuracy of ours is 60.14 and 62.12%, which are close to 60.3 and 64.1% of the basic SS-CNN and RSS-CNN presented in [10], respectively. We believe that the performance of our network could be further enhanced if we do the same post-processing as [10]. In terms of relative improvement in accuracy, the proposed DSAPN achieved a maximum increase of 4% on "*Safety*" and an average increase of 2% on all six attributes compared with RSS-CNN. Therefore, we consider the improvement is significant. Besides, the average performance of the DSAPN is enhanced by 0.65% than that of the DCPN, which indicates that semantic recognition task can improve the performance of urban visual perception tasks; that is to say, the semantic information of images contributes to visual urban perception. After that, we fine-tune the Block-4 and Block-5 in DCPN and DSAPN with the batch size of 32. Obviously, the performance of DSAPN (fine-tuned) and DCPN (fine-tuned) is significantly better than DSAPN and PRN. We analyze and conclude that the deep layer can better represent the abstract features of the data than shallow layer. So the fine-tuned networks have higher correlation with the training target and achieve better performance. Additionally, DSAPN (fine-tuned) still has excellent performance compared to state-of-the-art RSS-CNN and SS-CNN. There is no doubt that DSAPN (fine-tuned) achieves the best performance on six attributes for visual urban perception, demonstrating the effectiveness of our method.

Table 1 Results of visual urban perception prediction

Attribute	DCPN		DSAPN		SS-CNN	RSS-CNN
	ResNet	Fine-tune	ResNet	Fine-tune		
Safety	59.76	60.63	62.95	**64.87**	60.14	62.12
Lively	58.76	59.90	59.42	**61.09**	60.05	59.95
Beautiful	67.83	68.79	68.30	**69.20**	68.67	69.12
Wealthy	65.44	65.82	64.30	**65.91**	65.53	65.35
Boring	56.32	56.47	56.40	**57.47**	56.34	56.04
Depressing	60.15	61.26	60.80	**62.10**	60.89	61.29

5.3 Zero-shot learning for newly-added attributes

5.3.1 Semantic correlation analysis

To further demonstrate the effectiveness of semantic information and investigate how semantic information influences urban perception prediction task again, we analyse the correlation among them. The semantic correlation matrix $\prod$ is shown in Fig. 5, which is obtained by 500 iterations on six perceptual attributes using pre-trained DSAPN.

As we can see from the Fig. 5, there is a high correlation between the perceptual attributes and specific semantic information, where blue represents positive correlation, and red represents negative correlation. The darker the color, the stronger the correlation. For example, the "*prison*" in the image is positively related to the attribute "*Depressing*", but negatively related to "*Safety*", which is consistent with our common sense. In our daily life, prisons often leave an impression of oppression and insecurity, while natural landscapes often give people a lively feeling. The result proves once again that semantic information does have positive significance for visual urban perception task.

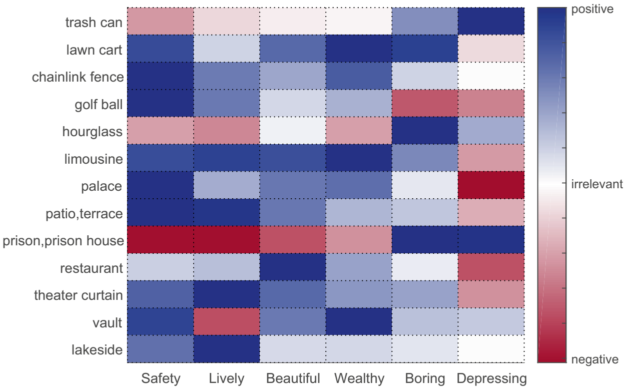

Fig. 5 The relationship between semantic information and perceptual attributes. The significance of the correlation is illustrated by the color depth in a heat map manner. Blue represents positive correlation and red represents negative correlation. The darker the color, the stronger the correlation (colour figure online)

5.3.2 Cross-attributes prediction

From the experimental results mentioned above, we can find that tall buildings contained in images are perceived as "*Safety*" and "*Wealthy*", respectively. This discovery makes us wonder that there may be a connection between different attributes. To further explore the relationships between six perceptual attributes, we train a separate DSAPN model for each perceptual attribute and evaluate the performance of each model on the specific perception attribute. For instance, the DSAPN for "*Safety*" is trained only with label of "*Safety*" of training images for visual urban perception task. The test sample is fixed when predicting the accuracy of six perceptual attributes on each DSAPN model. The results are shown in Table 2.

The experimental results indicate that the models trained by "*Safety*", "*Lively*", "*Beautiful*" and "*Wealthy*" separately have excellent performance on the other three perceptual attributes, while the performance on the "*Boring*" and "*Depressing*" attributes is disappointed. Similarly, the models trained by "*Boring*" and "*Depressing*" separately have delightful performance between each other and get poor performance on "*Safety*", "*Lively*", "*Beautiful*" and "*Wealthy*". The reason for this result may be that "*Safety*", "*Lively*", "*Beautiful*" and "*Wealthy*" are attributes with positive meaning, so they are more similar in semantic expression. But "*Boring*" and "*Depressing*" are attributes with negative meaning, so they are more similar in semantic expression. Besides, there is a large semantic gap between these two types of positive and negative attributes. It can be concluded that there is a strong correlation between different perceptual attributes.

5.3.3 Zero-shot prediction for newly-added attributes

In this section, we leverage the deep semantic-aware network for zero-shot (DSANZS) model to predict perception score of testing images on newly-added attributes. Specifically, in each zero-shot learning architecture, we randomly choose five attributes as training attributes, the remaining one as the newly-added attribute. We first train the DSAPN model based on training attribute and obtain

Table 2 Cross-attribute prediction performance

Test	Train					
	Safety	Lively	Beautiful	Wealthy	Boring	Depressing
Safety	**64.87**	57.78	57.09	58.94	47.42	42.92
Lively	59.05	**61.09**	56.36	59.17	43.24	41.22
Beautiful	64.06	58.28	**69.20**	63.97	48.08	34.83
Wealthy	63.67	62.66	61.78	**65.91**	41.75	36.38
Boring	45.24	42.11	47.60	44.52	**57.47**	54.52
Depressing	39.65	40.01	39.65	39.15	56.41	**62.10**

Table 3 Zero-shot learning accuracy

Attribute	DSANZS	NN	ConSE	ConSE-pseudo
Safety	**56.71**	55.14	55.32	55.81
Lively	**55.42**	55.31	54.93	55.29
Beautiful	**57.88**	56.11	56.77	56.94
Wealthy	**58.41**	57.79	58.01	58.32
Boring	**55.41**	54.98	54.79	55.22
Depressing	**56.22**	55.11	55.97	56.13

the SCM matrix. Then based on the obtained SCM matric, we use the defined unseen attribute and testing image representation to do visual urban perception prediction. Table 3 shows the accuracy results of different zero-shot learning methods for visual urban perception task on PP 2.0 dataset.

To show the effectiveness of our proposed method, we compare it with the following methods:

- **Nearest neighbor (NN)** uses Eq. (7) to synthesize the prototypes of zero-shot attributes and utilize $\boldsymbol{F}_{\hat{z}}(I) = cos(\mathbf{x}_i^S, \prod_{\hat{z}})$ to get the perceptual score for image I_i.
- **ConSE** uses the same $p(\cdot)$ function to predict the posterior of one testing image belonging to each known perceptual attribute. It tries to synthesize the testing-instance prototype using Eq. (8) by replacing the semantic representations with 1000-dimensional word vectors for each attribute. The zero-shot attribute prototype, demonstrating in Eq. (2), is also replaced by a 1000-dimensional word vector.
- **ConSE-pseudo** is a variant of ConSE that is more comparable to our method. The main difference between ConSE-pseudo and our method is that ConSE-pseudo replaces our semantic representation of known and unknown attributes with a 1000-dimensional word vector for zero-shot recognition in both Eqs. (7) and (8).

From Table 3, we can see that, our DSANZS model is better than all the other baselines on the same dataset. It achieves better performance than ConSE and ConSE-pseudo, which are the state-of-the-art approaches for zero-shot learning. One of the most important reasons for our progress is that we obtain the semantic representation by mining correlations between perceptual attributes and semantic information of images, which is much semantically discriminative than word vectors trained by text corpus. The results illustrate that using the semantic representation of images can offer better zero-shot prediction performance on visual urban perception.

6 Conclusions and future work

In this paper, we explore the issue of visual urban perception from a new perspective. We define the visual urban perception problem as a two step zero-shot learning problem, which includes the supervised visual urban perception step for training attributes and the zero-shot prediction step for newly-added attributes. In the first step, we develop a novel network structure and introduce the semantic information into supervised visual urban perception. In the second step, we use the visualization techniques to obtain the semantic correlation between objects and training attributes from the well trained supervised models. Based on the semantic correlation space, we obtain the representation of unseen attribute and testing image, and further predict the perception scores for images on unseen perceptual attributes. Experimental results show that our supervised DSAPN model and zero-shot prediction on newly-added attributes are superior to their corresponding state-of-the-art methods, which also implicitly reflects that the new two step zero-shot learning architecture of our proposed DSANZS is effective. Besides, we agree that the styles of different cities and countries may be concerned with visual urban perception, and we will investigate it in our future work.

Acknowledgements This work is supported by the National Natural Science Foundation of China (62077033, 61876098, 61976123), Shandong Provincial Natural Science Foundation Key Project (ZR2020KF015), Major Scientific and Technological Innovation Projects of Shandong Province (2018CXGC1501), the Fostering Project of Dominant Discipline and Talent Team of Shandong Province Higher Education Institutions, and the Taishan Young Scholars Program of Shandong Province and Key Development Program for Basic Research of Shandong Province (ZR2020ZD44).

References

1. Akata Z, Reed S, Walter D, Lee H, Schiele B (2015) Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2927–2936
2. Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. J Mach Learn Res 3(Feb):1137–1155
3. Can G, Benkhedda Y, Gatica-Perez D (2018) Ambiance in social media venues: Visual cue interpretation by machines and crowds. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 2363–2372
4. Chollet F et al (2015) Keras
5. Cohen DA, Mason K, Bedimo A, Scribner R, Basolo V, Farley TA (2003) Neighborhood physical conditions and health. Am J Public Health 93(3):467–471
6. David HA (1960) The method of paired comparisons. In: Proceedings of the fifth conference on the design of experiments in army research developments and testing, pp 1–16

7. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv:1810.04805
8. Deza A, Parikh D (2015) Understanding image virality. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1818–1826
9. Dosovitskiy A, Brox T (2016) Inverting convolutional networks with convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4829–4837
10. Dubey A, Naik N, Parikh D, Raskar R, Hidalgo CA (2016) Deep learning the city: quantifying urban perception at a global scale. In: European conference on computer vision. Springer, Berlin, pp 196–212
11. Farhadi A, Endres I, Hoiem D, Forsyth D (2009) Describing objects by their attributes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, pp 1778–1785
12. Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Mikolov T et al (2013) Devise: a deep visual-semantic embedding model. In: Proceedings of neural information processing systems, pp 2121–2129
13. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
14. He S, Yoshimura Y, Helfer J, Hack G, Ratti C, Nagakura T (2020) Quantifying memories: mapping urban perception. Mob Networks Appl 2020(25):1275–1286
15. Hu CB, Zhang F, Gong FY, Ratti C, Li X (2020) Classification and mapping of urban canyon geometry using google street view images and deep multitask learning. Build Environ 167:106424
16. Isola P, Xiao J, Torralba A, Oliva A (2011) What makes an image memorable? In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 145–152
17. Jayasuriya M, Arukgoda J, Ranasinghe R, Dissanayake G (2020) Localising PMDs through CNN based perception of urban streets. In: Proceedings of the IEEE international conference on robotics and automation (ICRA). IEEE, pp 6454–6460
18. Jeon JY, Jo HI (2020) Effects of audio–visual interactions on soundscape and landscape perception and their influence on satisfaction with the urban environment. Build Environ 169:106544
19. Jiang H, Wang R, Shan S, Chen X (2019) Transferable contrastive network for generalized zero-shot learning. In: Proceedings of the IEEE international conference on computer vision, pp 9765–9774
20. Kao Y, He R, Huang K (2017) Deep aesthetic quality assessment with semantic information. IEEE Trans Image Process 26(3):1482–1495
21. Koniusz P, Yan F, Mikolajczyk K (2013) Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection. Comput Vis Image Underst 117(5):479–492
22. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
23. Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 951–958
24. Larochelle H, Erhan D, Bengio Y (2008) Zero-data learning of new tasks. AAAI 1:3
25. Law S, Paige B, Russell C (2019) Take a look around: using street view and satellite images to estimate house prices. ACM Trans Intel Sys Technol (TIST) 10(5):1–19
26. Li J, Jing M, Lu K, Ding Z, Zhu L, Huang Z (2019) Leveraging the invariant side of generative zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7402–7411
27. Li K, Min MR, Fu Y (2019) Rethinking zero-shot learning: a conditional visual classification perspective. In: Proceedings of the IEEE international conference on computer vision, pp 3583–3592
28. Liu L, Zhang H, Xu X, Zhang Z, Yan S (2019) Collocating clothes with generative adversarial networks cosupervised by categories and attributes: a multidiscriminator framework. IEEE Trans Neural Netw Learn Syst 31(9):3540–3554
29. Liu M, Zhang D, Chen S (2014) Attribute relation learning for zero-shot classification. Neurocomputing 139:34–46
30. Liu X, Chen Q, Zhu L, Xu Y, Lin L (2017) Place-centric visual urban perception with deep multi-instance regression. In: Proceedings of the ACM on multimedia conference, pp 19–27
31. Mahendran A, Vedaldi A (2015) Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5188–5196
32. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint. arXiv:1301.3781
33. Milam A, Furr-Holden C, Leaf P (2010) Perceived school and neighborhood safety, neighborhood violence and academic achievement in urban school children. Urban Rev 42(5):458–467
34. Min W, Mei S, Liu L, Wang Y, Jiang S (2019) Multi-task deep relative attribute learning for visual urban perception. IEEE Trans Image Process 29:657–669
35. Naik N, Philipoom J, Raskar R, Hidalgo C (2014) Streetscore-predicting the perceived safety of one million streetscapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 779–785
36. Nasar JL (1990) The evaluative image of the city. J Am Plan Assoc 56(1):41–53
37. Norouzi M, Mikolov T, Bengio S, Singer Y, Shlens J, Frome A, Corrado GS, Dean J (2013) Zero-shot learning by convex combination of semantic embeddings. arXiv preprint. arXiv:1312.5650
38. Ordonez V, Berg TL (2014) Learning high-level judgments of urban perception. In: Proceedings of the European conference on computer vision. Springer, pp 494–510
39. Piro FN, Nœss Ø, Claussen B (2006) Physical activity among elderly people in a city population: the influence of neighbourhood level violence and self perceived safety. J Epidemiol Community Health 60(7):626–632
40. Porzi L, Rota Bulò S, Lepri B, Ricci E (2015) Predicting and understanding urban perception with convolutional neural networks. In: Proceedings of the 23rd ACM international conference on Multimedia. ACM, pp 139–148
41. Qiao R, Liu L, Shen C, Van Den Hengel A (2016) Less is more: zero-shot learning from online textual documents with noise suppression. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2249–2257
42. Quercia D, O'Hare NK, Cramer H (2014) Aesthetic capital: what makes London look beautiful, quiet, and happy? In: Proceedings of the 17th ACM conference on computer supported cooperative work & social computing, pp 945–955
43. Rohrbach M, Stark M, Szarvas G, Gurevych I, Schiele B (2010) What helps where—and why? Semantic relatedness for knowledge transfer. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 910–917
44. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
45. Salesses P, Schechtner K, Hidalgo CA (2013) The collaborative image of the city: mapping the inequality of urban perception. PLoS One 8(7):e68400
46. Sariyildiz MB, Cinbis RG (2019) Gradient matching generative networks for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2168–2178

47. Shen Y, Qin J, Huang L, Liu L, Zhu F, Shao L (2020) Invertible zero-shot recognition flows. In: Proceedings of European conference on computer vision. Springer, Berlin, pp 614–631
48. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint. arXiv:1312.6034
49. Sistu G, Leang I, Chennupati S, Hughes C, Milz S, Yogamani S, Rawashdeh S (2019) NeurAll: towards a unified model for visual perception in automated driving. arXiv preprint. arXiv:1902.03589
50. Socher R, Ganjoo M, Manning CD, Ng A (2013) Zero-shot learning through cross-modal transfer. In: Advances in neural information processing systems, pp 935–943
51. Tenney I, Das D, Pavlick E (2019) Bert rediscovers the classical nlp pipeline. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 4593–4601
52. Wang Q, Chen K (2017) Alternative semantic representations for zero-shot human action recognition. In: Proceedings of the Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, pp 87–102
53. Wang W, Zheng VW, Yu H, Miao C (2019) A survey of zero-shot learning: settings, methods, and applications. ACM Trans Intell Syst Technol 10(2):1–37
54. Wilson JQ (2003) Broken windows: the police and neighborhood safety. In: Proceedings of the social, ecological and environmental theories of crime. Routledge, pp 169–178
55. Wu Z, Fu Y, Jiang YG, Sigal L (2016) Harnessing object and scene semantics for large-scale video understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3112–3121
56. Xian Y, Akata Z, Sharma G, Nguyen Q, Hein M, Schiele B (2016) Latent embeddings for zero-shot classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 69–77
57. Xie GS, Liu L, Jin X, Zhu F, Zhang Z, Qin J, Yao Y, Shao L (2019) Attentive region embedding network for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9384–9393
58. Xie GS, Liu L, Zhu F, Zhao F, Zhang Z, Yao Y, Qin J, Shao L (2020) Region graph embedding network for zero-shot learning. In: Proceedings of the European conference on computer vision. Springer, Berlin, pp 562–580
59. Xu Y, Yang Q, Cui C, Shi C, Song G, Han X, Yin Y (2019) Visual urban perception with deep semantic-aware network. In: International conference on multimedia modeling. Springer, Berlin, pp 28–40
60. Yao Y, Liang Z, Yuan Z, Liu P, Bie Y, Zhang J, Wang R, Wang J, Guan Q (2019) A human–machine adversarial scoring framework for urban perception assessment using street-view images. Int J Geogr Inf Sci 33(12):2363–2384
61. Zhang F, Zhou B, Liu L, Liu Y, Fung HH, Lin H, Ratti C (2018) Measuring human perceptions of a large-scale urban region using machine learning. Landsc Urban Plan 180:148–160
62. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929
63. Zhou B, Liu L, Oliva A, Torralba A (2014) Recognizing city identity via attribute analysis of geo-tagged images. In: Proceedings of the European conference on computer vision. Springer, Berlin, pp 519–534