sustainability

MDPI

Article

Semantic and Instance Segmentation in Coastal Urban Spatial Perception: A Multi-Task Learning Framework with an Attention Mechanism

Hanwen Zhang, Hongyan Liu and Chulsoo Kim *

Department of Marine Design Convergence Engineering, Pukyong National University, 45, Yongso-ro, Nam-Gu, Busan 48513, Republic of Korea; zhanghanwen@pukyong.ac.kr (H.Z.); liuyongyan@pukyong.ac.kr (H.L.)
* Correspondence: kimcsoo@pknu.ac.kr; Tel.: +82-051-629-5361

Abstract: With the continuous acceleration of urbanization, urban planning and design require more in-depth research and development. Street view images can express rich urban features and guide residents' emotions toward a city, thereby providing the most intuitive reflection of their perception of the city's spatial quality. However, current researchers mainly conduct research on urban spatial quality through subjective experiential judgment, which includes problems such as a high cost and a low judgment accuracy. In response to these problems, this study proposes a multi-task learning urban spatial attribute perception model that integrates an attention mechanism. Via this model, the existing attributes of urban street scenes are analyzed. Then, the model is improved by introducing semantic segmentation and instance segmentation to identify and match the qualities of the urban space. The experimental results show that the multi-task learning urban spatial attribute perception model with an integrated attention mechanism has prediction accuracies of 79.54%, 78.62%, 79.68%, 77.42%, 78.45%, and 76.98% for the urban spatial attributes of beauty, boredom, depression, liveliness, safety, and richness, respectively. The accuracy of the multi-task learning urban spatial scene feature image segmentation model with an integrated attention mechanism is 95.4, 94.8, 96.2, 92.1, and 96.7 for roads, walls, sky, vehicles, and buildings, respectively. The multi-task learning urban spatial scene feature image segmentation model with an integrated attention mechanism has a higher recognition accuracy for urban spatial buildings than other models. These research results indicate the model's effectiveness in matching urban spatial quality with public perception.

Keywords: machine learning; attention mechanism; urban spatial quality; multi-task learning; public perception

Citation: Zhang, H.; Liu, H.; Kim, C. Semantic and Instance Segmentation in Coastal Urban Spatial Perception: A Multi-Task Learning Framework with an Attention Mechanism. *Sustainability* **2024**, *16*, 833. https://doi.org/10.3390/su16020833

Academic Editors: Wen-Hsien Tsai, Ahmad Hassanat, Sami Mnasri and Ahmad S. Tarawneh

Received: 26 October 2023
Revised: 25 December 2023
Accepted: 28 December 2023
Published: 18 January 2024

1. Introduction

With economic development and the continuous acceleration of urbanization, humanism has become a core concept within urban planning and design [1]. The study of urban spatial quality is an important aspect of improving street living environments, and each city has different developmental stages and functional positionings. Urban residents are more inclined toward beautiful living environments, convenient travel, and safe road facilities, while urban tourists and mobile populations are more inclined toward prosperous urban streets and rich and colorful areas [2]. The evaluation of urban space requires support from precise and scientific evaluation systems. With the emergence of many urban street-view images, images have played an important role in quantifying urban spatial quality and the public's preference for urban quality. This study proposes a multi-task learning urban spatial attribute perception model that integrates an attention mechanism to quantify urban spatial attributes. This model can extract and match street attributes. We then improve the model to address the problem of global and partial extraction in urban street images. A multi-task learning urban spatial scene element image segmentation model with an

integrated attention mechanism is proposed, which can analyze the correlation between urban scene elements and urban spatial perception attributes via quantitative and qualitative analyses. Examples include green space and resident satisfaction, cultural facilities, and cultural attractions. This study aims to provide academic support for urban spatial quality research and a major guide for making judgements within urban planning and design. In addition, it aims to provide a precise understanding of residents' and tourists' expectations and needs for urban spaces to better meet these expectations and enhance the urban living and visiting experience. It also aims to enable a more objective and comprehensive assessment of the quality of urban space, rather than relying solely on subjective opinions or small-scale surveys, making urban planning more responsive to future needs and promoting the development of cities in a more attractive and sustainable direction. The needs of residents and tourists can be better met by analyzing their intended characteristics and perceived preferences for urban cultural spaces. Through the optimization of cultural space, a design that meets the public's preferences can become a unique symbol of the city, improve the city's competitiveness, attract more people, and promote the development of cultural tourism and related industries. This research is divided into four main parts: The first part briefly describes other scholars' research on urban cultural space. The second part explains the attention mechanism and multi-task learning and establishes the multi-task learning urban spatial attribute perception model integrated with an attention mechanism via two methods. Then, the model is improved, and the multi-task learning urban spatial scene element image segmentation model integrated with an attention mechanism is established. In the third part, MTA-PM's accuracy under different iteration numbers and its prediction accuracy for different street attributes is compared with that of AlexNet, ResNet, and GoogleNet. Then, MTA-SM's accuracy is compared with that of other methods with different numbers of iterations and datasets, and randomly selected people evaluate the model. The fourth part summarizes all the above studies and discusses prospects for future research.

2. Domestic and Foreign Research

In the process of urbanization, humanism has become a core concept within urban planning and design. Wang R. et al. [3] found that most researchers had conducted research on aesthetic preference or stress release in urban green spaces, but few had conducted research on both issues simultaneously. In response to this issue, the research team proposed a direct scoring method based on a single measurement, which can analyze aesthetic preference and stress release in relation to urban environmental characteristics. The experimental results indicated that individual measurements can be used to evaluate the resilience of perception. The potential applications of urban green space practical design include planting more green plants [3]. Their study provides research directions for this research on the intentional characteristics of urban cultural spaces and the public's perceived preferences. In addition, Wang D. et al. found that traditional urban planning required a significant amount of time due to building constraints. In response to this issue, their research team proposed a cascaded deep generative adversarial network to plan urban space based on people's needs. The results showed that it can perform urban planning for complex urban environments and demonstrate reasonable logic [4]. The cascaded deep generative adversarial network (cascaded D-GAN) is a sophisticated machine learning architecture used for generating realistic and high-quality urban images. It operates in a cascaded manner, where multiple generative adversarial networks (GANs) are stacked sequentially to progressively refine the output. Each stage of the cascade focuses on capturing different levels of detail and complexity in the generated images. The cascaded D-GAN provides a powerful tool for urban planners to visualize and explore various design possibilities. It can generate realistic images of urban spaces, enabling planners to assess the visual impacts of different architectural and layout choices. Lekus H. Y. discovered that the humanization of urban public spaces was an important strategic component in the process of urbanization. Therefore, a method for humanized urban space planning

was proposed. The experimental results showed that this method can accurately identify humanized designs in urban planning [5].

Hao Y. et al. found that the icing phenomenon in urban power lines can endanger the safety of circuits. Currently, ice monitoring for power lines relies on manual monitoring. In response to this issue, the research team proposed a weakly supervised and phased transfer learning method to identify whether circuits were icing. The results showed that it can effectively identify frozen lines, and its recognition speed was fast [6]. Liang J. et al. found that creating sufficient multi-label image recognition technology was a major challenge in the multimedia field. Although graph convolutional networks can effectively learn global images, they were rarely applied to local images. In response to this issue, the research team proposed a multi-scale semantic attention model, which included a multi-scale module, a semantic guided attention module, and a graph convolutional network module. The results showed that the model achieved classification accuracies of 83.4% and 94.2% on the MS-COO and PASCAL VOC2007 datasets, respectively, and demonstrated good performances [7]. Lou G. et al. found it was difficult to associate low-level information in images with high-level image semantics in image recognition. In response to this issue, the research team proposed a model using VGG16. This model can extract image features from low-level images and convert them into feature languages. The experimental results showed that the proposed model had a high recognition accuracy for each image [8].

In summary, many scholars have conducted research in the field of recognition and detection and urban spatial quality and have achieved significant results. However, most scholars have not combined the two methods to study the intention and characteristics of urban cultural space and perceive the public's preferences. This study proposes a multi-task learning urban spatial attribute perception model that integrates an attention mechanism. This model can extract and match the attributes of streets. In addition, this study proposes a multi-task learning urban spatial scene element image segmentation model that integrates an attention mechanism. This model can analyze the correlation between urban scene elements and urban spatial perception attributes via quantitative and qualitative analyses to meet the needs of urban residents and future urban planning and design.

3. Research Methodology

The first section of this chapter explains the attention mechanism and multi-task learning and establishes the multi-task learning urban spatial attribute perception model integrated with an attention mechanism. The second section improves the model and establishes a multi-task learning urban spatial scene element image segmentation model that integrates an attention mechanism. This model can analyze the correlation between urban scene elements and urban spatial perception attributes via quantitative and qualitative analysis.

3.1. Research on Multi-Task Learning Urban Spatial Quality Attribute Perception Model Integrating Attention Mechanism

Multi-task learning refers to the ability to share the weights of certain characteristics during the process of shared learning among multiple tasks. Multi-task learning has a better generalization ability than single-task learning. At present, the majority of machine learning is single-task learning, such as image classification (e.g., recognizing if an image contains a cat or a dog). However, many problems cannot be divided into a single problem. Subdividing a problem into subproblems may cause information loss between the subproblems [9]. Figure 1 shows the structures of single-task and multi-task learning.

In Figure 1, multi-task learning is based on shared representations, which can simultaneously learn multiple related tasks, such as the automated driving of cars. It is a derivation of the transfer learning method. Multi-task processing learning refers to the process of learning how to simultaneously execute multiple tasks by training a model. Traditional machine learning models can usually only solve a single task, while multi-task processing learning enables the model to simultaneously process multiple related tasks, thereby improving the model's efficiency and generalization ability. Each task has an independent

model for training [10]. However, the shared representation method for multi-task learning allows different tasks to share a portion of the learned feature representations by sharing the same representation layer in the model. Its advantage is that, when there are few training samples for a task, additional information can be provided through the training samples for other tasks, thereby improving the performance of the task. An attention mechanism refers to a technology that simulates human attention mechanisms in the fields of computer science and artificial intelligence, as shown in Figure 2.

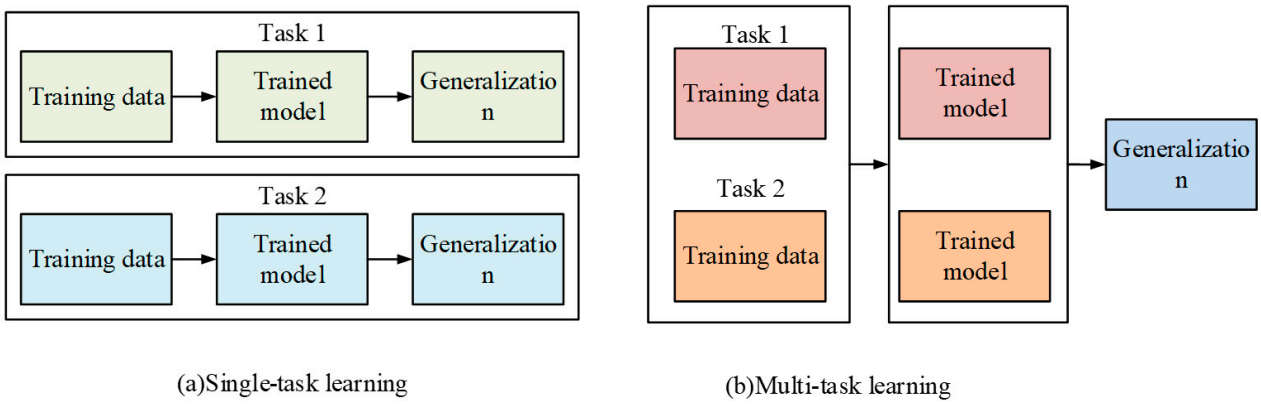

Figure 1. Single-task learning and multi-task learning structure diagrams.

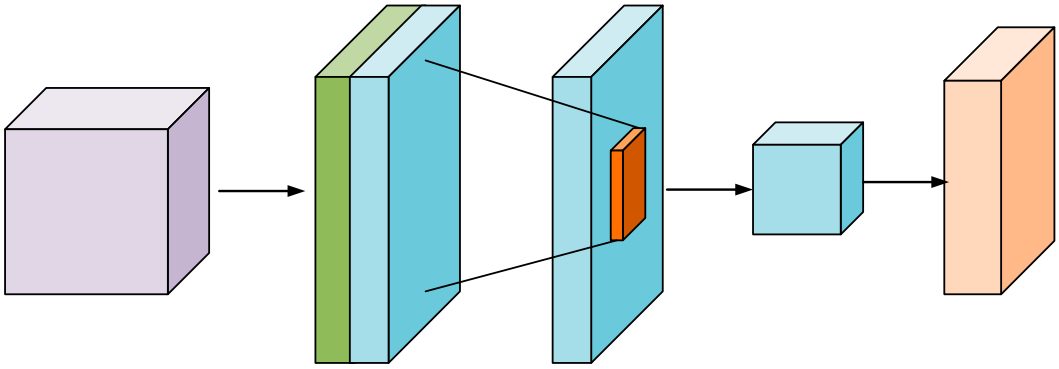

Figure 2. Schematic diagram of attention mechanism.

An attention mechanism improves the efficiency and accuracy of computers in processing large amounts of data and carrying out tasks by selectively focusing on and processing important information. The core idea of an attention mechanism is to enable computers to selectively focus on and process information related to the current task by learning and simulating human attention allocation methods while ignoring irrelevant information. This mechanism can help computers concentrate their energy, improve information processing, and reduce computing-resource waste. This study adopts the Visual Geometry Group Network (VGG-16) as the basic structure [11]. VGG-16 utilizes a relatively simple convolutional neural network architecture, making the entire network structure clear and easy to understand. This simplified architecture helps reduce the model's complexity, making it easier to train and tune. Moreover, the model achieves depth by stacking multiple convolutional layers, which are designed to help extract complex features from images. By downsampling multiple times, the network can gradually expand the sensory field to capture a wider range of information in the image. VGG-16 consists of multiple convolutional kernels and pooling layers, which can be divided into two parts, as shown in Figure 2.

According to Figure 3, the first part of VGG-16 contains two structures, each consisting of two 3 × 3 convolutional layers followed by a 2 × 2 max-pooling layer; the second part consists of three of those structures. Five layers of weight coefficients are used for feature

extraction and classification, with the extracted features eventually being aggregated into a fully connected layer [12]. Due to possible differences in data distribution between batches, data normalization is necessary to ensure a uniform distribution and avoid issues such as gradient vanishing, as shown in Equation (1).

$$y = \frac{x - mean(x)}{\sqrt{Var(x)} + eps} \times gamma + beta \quad (1)$$

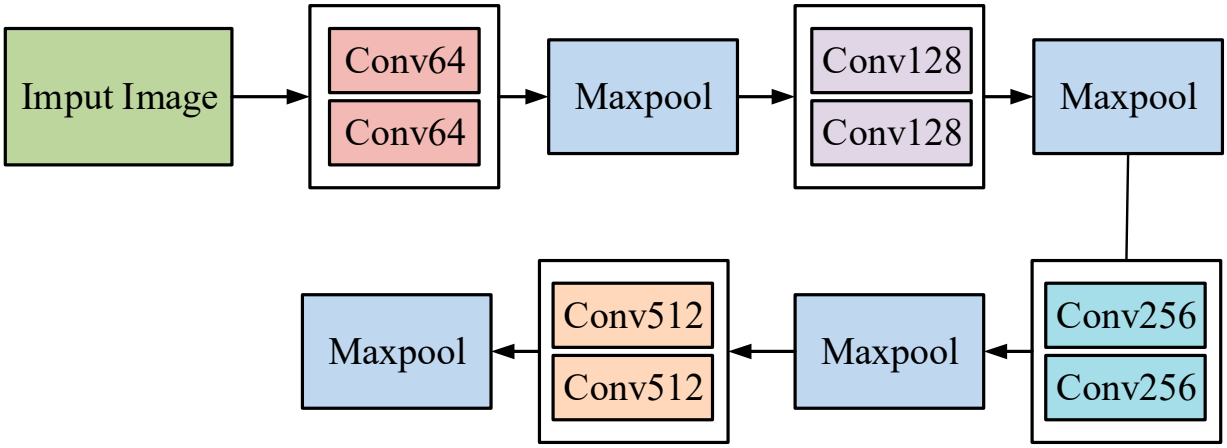

Figure 3. VGG-16's convolutional neural network framework.

In Equation (1), *gamma* and *beta* represent the learning parameters, and x represents the data input values. Images of roads in cities are learned via the multi-task learning twin convolutional neural network, VGG-16, and an attention model. The multi-task learning attention mechanism perception model for urban spatial quality attributes (MTA-PM) combines twin convolutional neural networks with VGG-16 and introduces an attention mechanism, which learns via classification and sorting subnetworks. It adopts an end-to-end network structure, making the attention mechanism and multi-tasking learning complement each other and work together.

The MTA-PM model process is as follows: First, images from two different datasets are input into the network, and then the images are simultaneously placed in the third dimension. The feature information in the images is extracted via the convolutional artificial neural network and attention mechanism to improve the target impact on attributes in specific attribute-learning tasks. The visual elements of the region are determined based on the learned attributes. Finally, feature matrix sparsity is used to determine the richness of attributes in different regional scenes. In practical scenarios, each scene type is interconnected, so multi-task learning can be used to establish a task with multiple connections. Via this task, parameters are shared during the feature extraction stage to distinguish and output them into the network. Considering the connections between different tasks, training can enhance imbalanced sample data by grouping the types of regions with connections. This study uses softmax classification to classify each image, as shown in Equation (2):

$$L_c = \sum_{(i,j,y)\in P} -[I[y=1]\log(g(x_i,x_j)) + [I[y=0]\log(1-g(x_i,x_j)) \quad (2)$$

In Equation (2), $I[\cdot]$ represents the indicator function, y represents the indicator variable, i and j represent images i and j, respectively, and $g(\cdot)$ represents the transformation

expression of twin convolutional neural networks. The ranking loss control framework ranks issues, as shown in Equation (3):

$$L_r = \min_{\omega} \sum_{(i,j,y)}^{N} [1 - y_{ij}\langle w, x_i - x_j \rangle]_+ + \lambda \|w\|^2 \quad (3)$$

In Equation (3), x_i represents the feature vector, λ represents the trade-off constant, and ω represents the parameter vector [13]. The multi-task loss function of the MTA-PM model is weighted and combined via sorting loss and classification loss, as shown in Equation (4):

$$L_{cr} = \min_{\omega} \sum_{p} L_c + \theta L_r \quad (4)$$

In Equation (4), L_c and L_r represent the classification loss and sorting loss. θ is obtained via network adaptive learning. The model can better cope with the convergence speed of different tasks by using the multi-task learning loss function to perform gradient descent [14]. In the field of urban spatial analysis, faster convergence of deep neural network training contributes to timelier and resource-efficient models. This, in turn, leads to more rapid decision making about spatial data in the urban planning and design process. The model improves the effect of the target on the attributes in an attribute-specific learning task by feeding the image into the network and then simultaneously placing the image in the third dimension to extract the feature information from the image via the convolutional artificial neural network and the attention mechanism. The learned attributes are used to determine the visual elements of the region, and, finally, the sparsity of the feature matrix is used to determine the richness of the attributes in the scenes of different regions.

3.2. Research on Multi-Task Learning Urban Spatial Scene Element Image Segmentation Model Integrating Attention Mechanism

Due to the fusion of attention mechanisms, multi-task learning models usually have many hyperparameters, including the weights of different tasks, the parameters of the attention mechanisms, etc. Tuning these hyperparameters may become complicated and require more experiments to find the optimal combination. The ability of the multi-task learning urban space and public perception preference models that integrate an attention mechanism to recognize urban space does not meet the requirements, so an image semantic segmentation model is introduced to improve the model [15]. Via semantic segmentation and instance segmentation, urban space can be divided into different semantic regions, such as roads, buildings, and green spaces. This helps to accurately recognize and understand the urban structure and provide detailed scene information for planning and design. Semantic segmentation makes it possible to analyze the distribution of different functional areas in the city in more detail, which allows us to better understand the city's spatial characteristics and functional layout. The semantic segmentation network (SegNet) for images is an improved image segmentation network using a fully convolutional neural network and VGG-16. Its structure is shown below.

In Figure 4, Seg's encoder extracts features from the image via the first 13 convolutional neural networks in VGG-16, reduces the image size, and increases the receptive field via the pooling layers [16]. The decoder in the SegNet model corresponds to the network layers in the encoder one by one, and then the image is restored to its original size via the collection layer. The final segmentation result is obtained via the classifier. In the convolution operation, inputting the feature map into the filter can obtain its output feature size, as shown in Equation (5).

$$h' = \left[\frac{h - f + s + p}{s}\right], w' = \left[\frac{w - f + s + p}{s}\right] \quad (5)$$

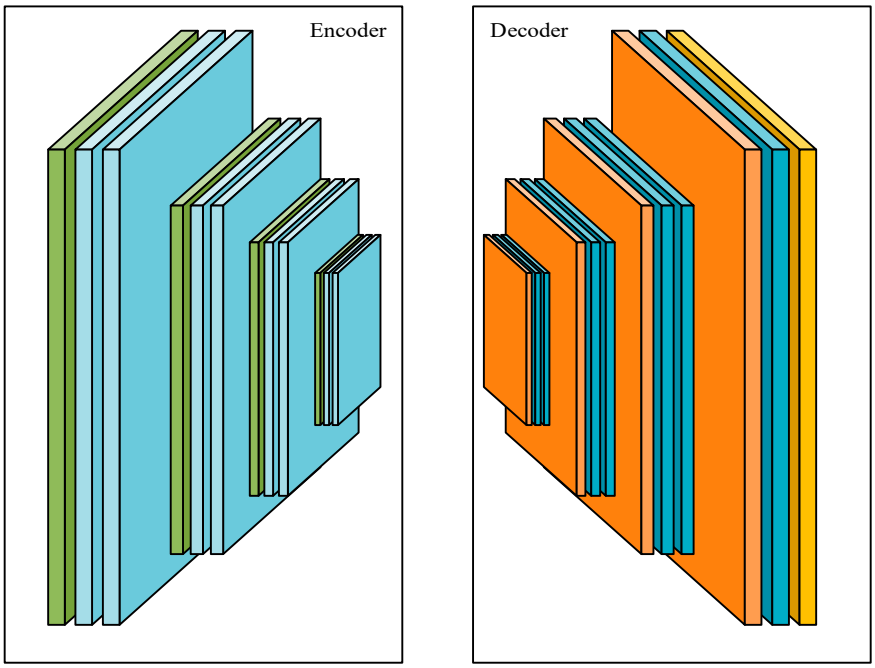

Figure 4. SegNet structure diagram.

In Equation (5), h' and w' represent the length and width of the output feature image, respectively, and h and w represent those of the input feature image; s represents the step size, and f represents the filter size. The loss function is shown in Equation (6):

$$L(p, y) = -\sum_{n} y_n \cdot \log(p_n), n \in [1, N] \tag{6}$$

In Equation (6), N is the neuron count, p represents the output type probability, and y represents the output. The multi-task learning attention mechanism image segmentation model for urban spatial scene elements (MTA-SM), which integrates an attention mechanism, uses a single image as the input instruction to generate more precise semantic segmentation. MTA-SM is composed of four parts, namely, the multi-task learning architecture, encoder, decoder, and attention module. The decoder and encoder select SegNet as the framework [17]. An attention mechanism is introduced into the specific convolutional layer of the decoder. Attention mechanisms are often used to enhance a model's focus on important regions, thereby improving the model's performance in a given task. In image segmentation tasks, the attention mechanism can be used to emphasize regions of interest and increase the model's focus on meaningful features. The attention module is connected to the shared network of VGG-16. By sharing networks to obtain information about the characteristics of local regions, each attention module includes a feature-processing method [18]. These attention modules can be used for specific tasks based on their characteristics. MTA-SM's structure is shown below.

As shown in Figure 5, each task has an independent weight-corresponding decoder, which introduces an attention module into the convolutional layer and utilizes the attention mask to rely on a shared feature network. Therefore, the features and attention masks in the shared networks can learn from each other, greatly improving the efficiency of multi-task learning.

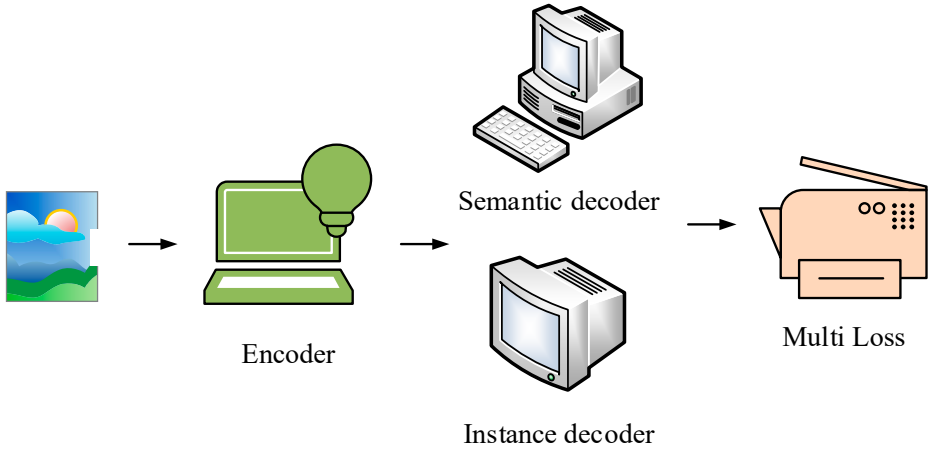

Figure 5. Multi-task learning image model incorporating attention mechanism.

The design of the attention module uses attention masks to extract features from the shared network. Attention masks can be used to indicate useless areas that are ignored by the attention mechanism, thereby improving the attention mechanism, limiting the attention range, and reducing unnecessary calculations and interference [19]. The task-specific features of the shared network are shown in Equation (7).

$$\hat{a}^{(i)} = p^{(i)} \odot a^{(i)} \tag{7}$$

In Equation (7), $p^{(i)}$ and $a^{(i)}$ represent convolutional layers with batch normalization, and $\odot$ represents element-wise multiplication. The structure of the encoder is shown in Figure 6.

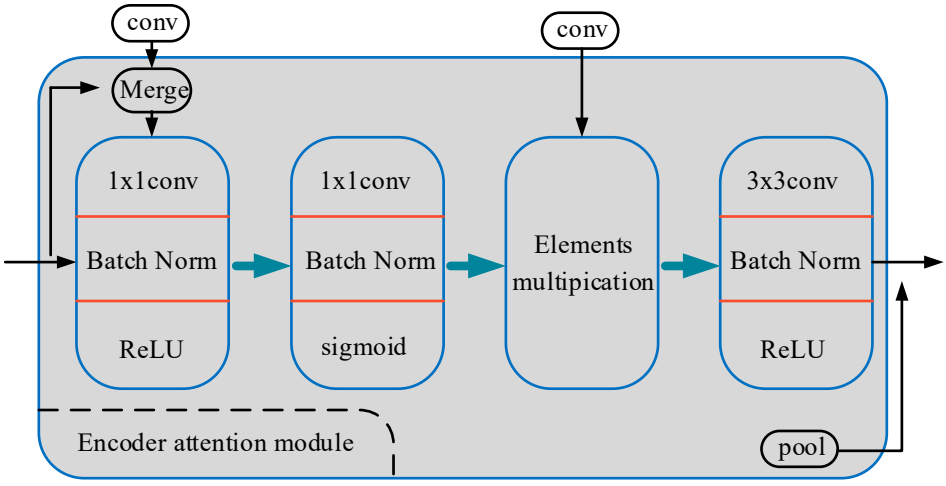

Figure 6. Encoder attention module.

As shown in Figure 6, the characteristics of the shared network are first input into the encoder's attention module. During the encoding process, the encoder processes each datum, and the second attention module only needs to extract relevant features from the shared network. Using this method can improve the encoder's performance. The input for the module's subsequent attention module is shown in Equation (8):

$$p^{(1)} = f^{(i)}(g^{(i)}(u^{(i)}; h^{(i)}(a^{(i)}))) \tag{8}$$

In Equation (8), $h^{(i)}$, $g^{(i)}$, and $f^{(i)}$ represent a batch-normalized convolutional layer that follows nonlinear activation. The feature extractor consists of two convolutions, which transmit the features to another attention module via the feature extractor and then match

the image resolution via the pooling layer. The loss function in multi-task learning is shown in Equation (9):

$$L = \sum_{i=1}^{k} \lambda_i L_i(X, Y) \tag{9}$$

In Equation (9), λ_i represents the task weight, X represents the input data, and Y represents the ground truth labels. For image segmentation, one datum represents all segmentation tasks in a domain. The loss function in semantic segmentation is shown in Equation (10):

$$L_1 = -\frac{1}{pq}\sum Y(p,q)\log \hat{Y}(p,q) \tag{10}$$

In Equation (10), $\hat{Y}$ represents the predicted value of the network, p represents the output class probability, and q represents the output. This study adopts a regression algorithm for instance segmentation, and its loss function is shown in Equation (11):

$$L_2 = \frac{1}{|N_j|}\sum_{Nj} \|x_n - \hat{x}_n\|_1 \tag{11}$$

In Equation (11), x_n represents the ground truth label, and N_j represents the labeled pixels. Multiple tasks must be balanced during training, so a dynamic weighted adaptive weighting method is introduced. The final loss function is shown in Equation (12):

$$L = \lambda_1 L_1 + \lambda_2 L_2 \tag{12}$$

In Equation (12), L_1 represents the semantic segmentation loss function, λ_1 represents the loss function weight in semantic segmentation, L_2 represents the instance segmentation loss function, and λ_2 represents the loss function weight in instance segmentation. The rate of change in losses for each task is calculated to balance the task weights. The weight calculation for each task is shown in Equation (13):

$$\begin{cases} \lambda_i(t) = \frac{Ke^{(w_i(t-1)/T)}}{\sum_m e^{(w_m(t-1)/T)}} \\ w_i(t-1) = \frac{L_i(t-1)}{L_i(t-2)} \end{cases} \tag{13}$$

In Equation (13), w_i represents the relative decline rate, T represents the weight coefficient, t represents the iteration coefficient, and K represents the weight balance coefficient.

4. Research Results

4.1. Multi-Task Learning Model Integrating Attention Mechanism Performance Analysis for Urban Space and Public Perception Preferences

This experiment used a server with a CPU (Intel®CoreTMi7-9700CPU@2.40GHz×8) and GPU (NVIDIA GeForce RTX 1060), and Windows 10 as the operating system. To evaluate the specific performance of MTA-PM, AlexNet, MTA-PM, ResNet, and GoogleNet were introduced. The results are shown in Figure 7.

Figure 7a shows a comparison of the loss curves for the four methods in the training set, while Figure 7b shows those in the validation set. As shown in the figure, the loss function values of the four methods in the training set gradually decreased with the increase in iteration times and tended to stabilize after four iterations. Among the four methods, the MAT-PE algorithm model had the smallest loss function value. In the validation set, the loss function values for the four methods decreased as the iterations increased, but the displayed loss function values were unstable and fluctuated. Among the four methods, the MAT-PE loss function values were at the minimum value. The results comparing AlexNet, MTA-PM, ResNet, and GoogleNet in the training set show that MAT-PE's minimum loss function value after reaching four iterations was 0.11, which was lower than those of the other algorithms. In the validation set, MAT-PE's minimum loss function value after

reaching five iterations was 0.12, which was lower than those of the other algorithms. This indicates that the proposed model had a good performance. To further verify MAT-PE's performance, Attn-VGG and Mtl-VGG were introduced for comparison with this model, and the results are shown in Figure 8.

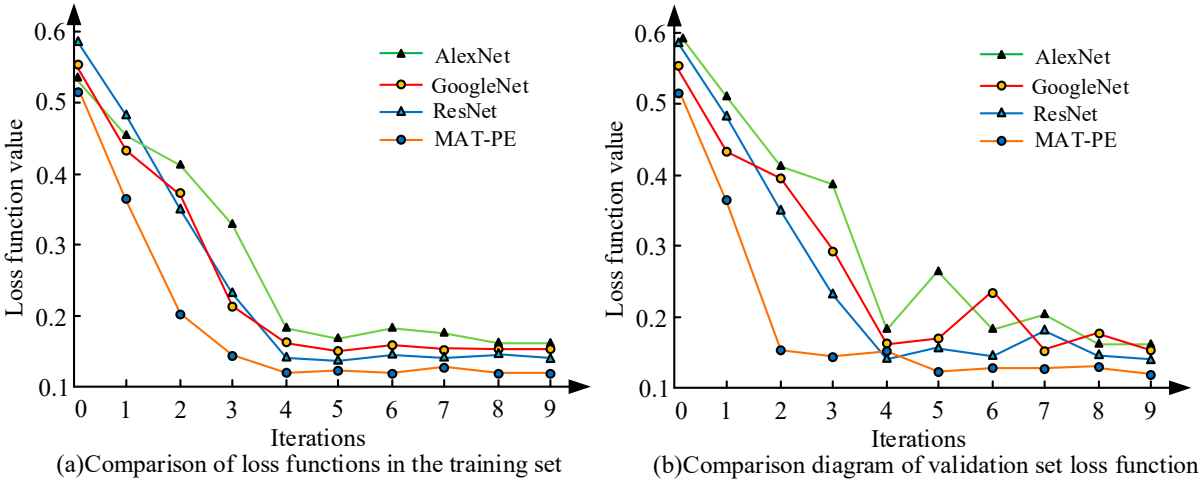

Figure 7. Comparison of loss curves in different datasets.

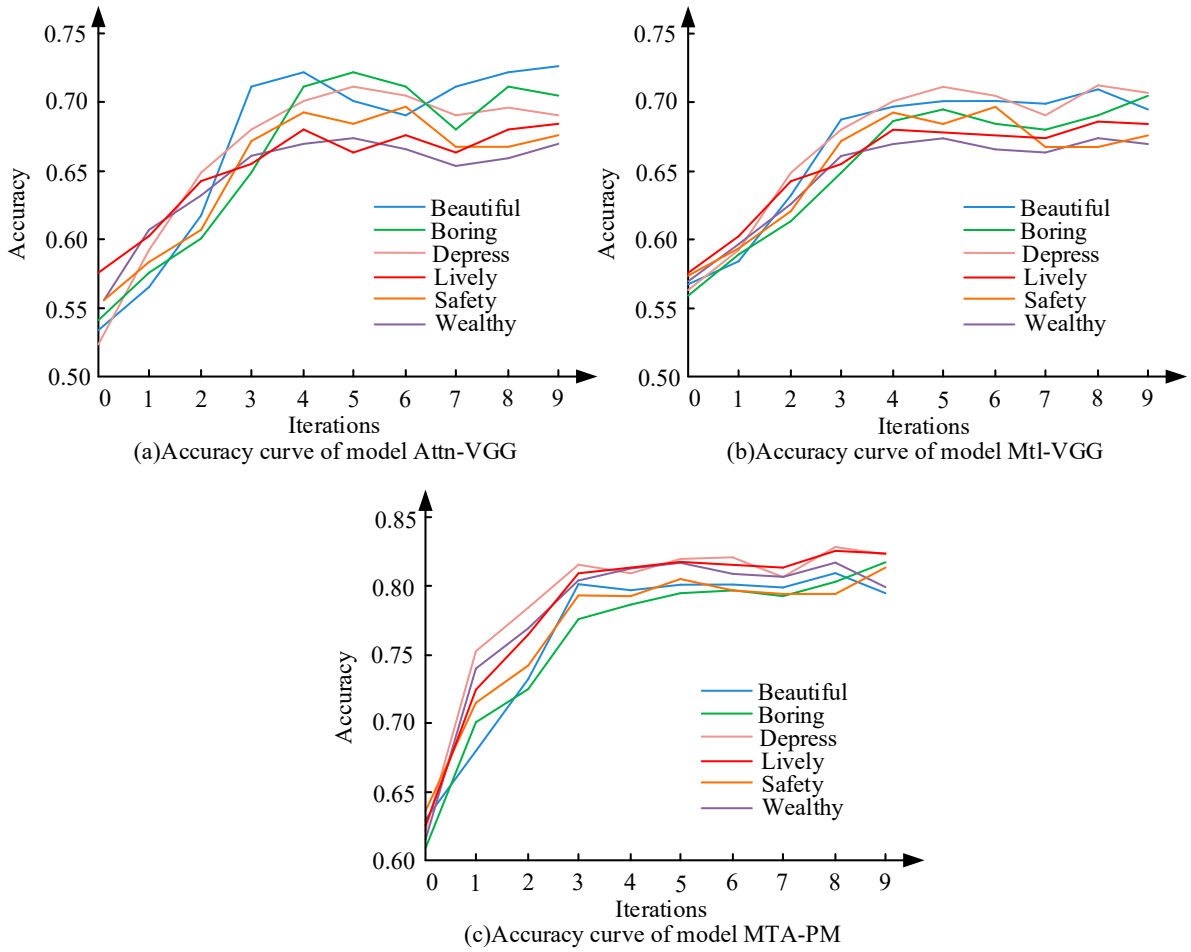

Figure 8. Accuracy of Attn-VGG, Mtl-VGG, and MTA-PM methods for different images.

Figure 8a shows the accuracy curve of the Attn-VGG model. The graph shows that, as the iterations increased, the accuracy increased accordingly. When the iterations reached

approximately four, the accuracy tended to reach its maximum. However, the accuracy was unstable and significantly fluctuated. Figure 8b shows the accuracy curve of the Mtl-VGG model. As the iterations increased, the model's accuracy increased accordingly. Moreover, the model's performance was more stable compared with the Attn-VGG model and tended to stabilize when the number of iterations was four. Figure 8c shows the accuracy curve of the MTA-PM model. As shown in the figure, the model achieved high accuracy when the number of iterations was three, and its performance was relatively stable with small fluctuations. The experimental results show that the proposed MTA-PM model can recognize the emotions expressed in different streets to a high level. Among the three methods, a better performance was achieved when the number of iterations was small, and the models' performances were relatively stable. Using the street attributes in the dataset that had been matched as experimental subjects, the three models were compared in the best performance scenario. The results are shown below.

As shown in Table 1, Vttn-VGG's prediction accuracies for the urban spatial attributes of beautiful, boring, depressing, lively, safe, and wealthy were 68.84%, 64.12%, 68.54%, 69.57%, 69.34%, and 68.31%, respectively. Mtl-VGG's prediction accuracies for the urban spatial public perception attributes of beautiful, boring, depressing, lively, safe, and wealthy were 70.84%, 70.96%, 70.65%, 71.65%, 71.74%, and 71.24%, respectively. MTA-PM's prediction accuracy rates for the urban spatial attributes of beautiful, boring, depressing, lively, safe, and wealthy were 79.54%, 78.62%, 79.68%, 77.42%, 78.45%, and 76.98%, respectively. The results show that the average prediction accuracies of Vttn-VGG, Mtl-VGG, and MTA-PM methods were 68.12%, 71.18%, and 78.45%. This indicates that the MTA-PM proposed in this study had the best performance.

Table 1. Accuracy of street-attribute prediction.

Accuracy (%)	Beautiful	Boring	Depressing	Lively	Safe	Wealthy	Average Value
Vttn-VGG	68.84	64.12	68.54	69.57	69.34	68.31	68.12
Mtl-VGG	70.84	70.96	70.65	71.65	71.74	71.24	71.18
MTA-PM	79.54	78.62	79.68	77.42	78.45	76.98	78.45

4.2. Urban Spatial Scene Element Segmentation Model Performance Analysis Integrating Attention Mechanism

To evaluate MTA-SM's performance, three different network architectures were designed, namely, SegNet, SegNet-AM, and Mtl-SegNet-AM. Comparative experiments were conducted with MTA-SM, and the results are shown in Figure 9.

Figure 9a shows the relationship between the two methods and iterations. MTP-SM reached approximately 30 iterations, and the model's matching accuracy stabilized. However, MTP-SM did not achieve the best model performance. MTP-SM's matching accuracy reached its maximum value when the number of iterations reached approximately 40. Figure 9b shows the relationship between the two methods and the training set size. As the training set size increased, the matching accuracy of the models also increased. When the model's training set size reached approximately 300, the changes in the accuracy of the two models tended to stabilize and achieve the best performance. The results show that MTP-SM can perform well with fewer iterations and a smaller training set. The different components of the city were recognized. The ratio of the number of recognized images to the total number of images was used as the recognition accuracy [20]. The recognition results are shown in Figure 10.

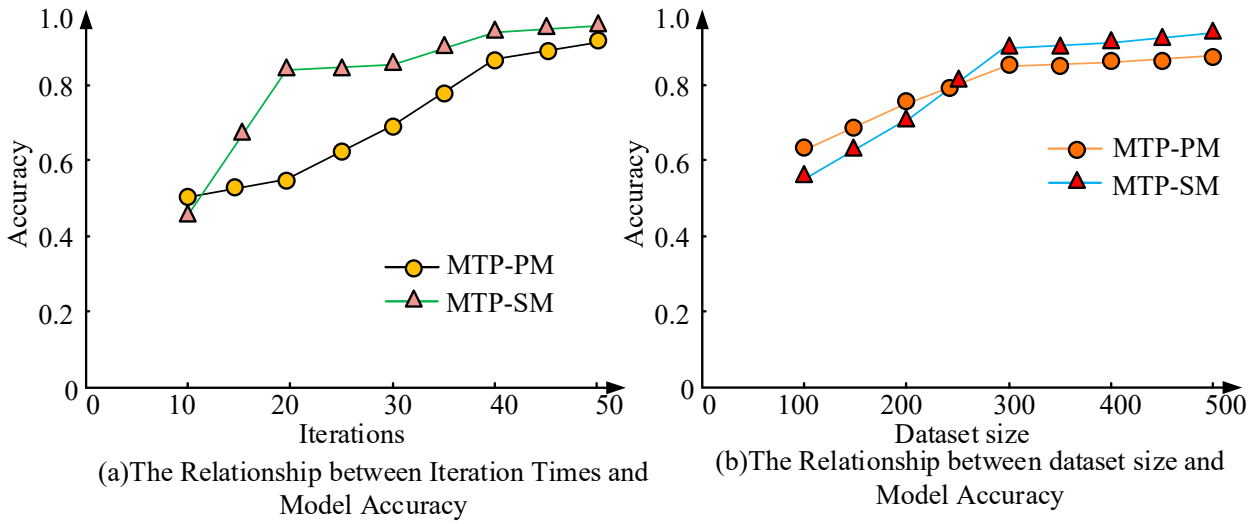

Figure 9. Comparison of iteration times, dataset size, and model matching accuracy between two models.

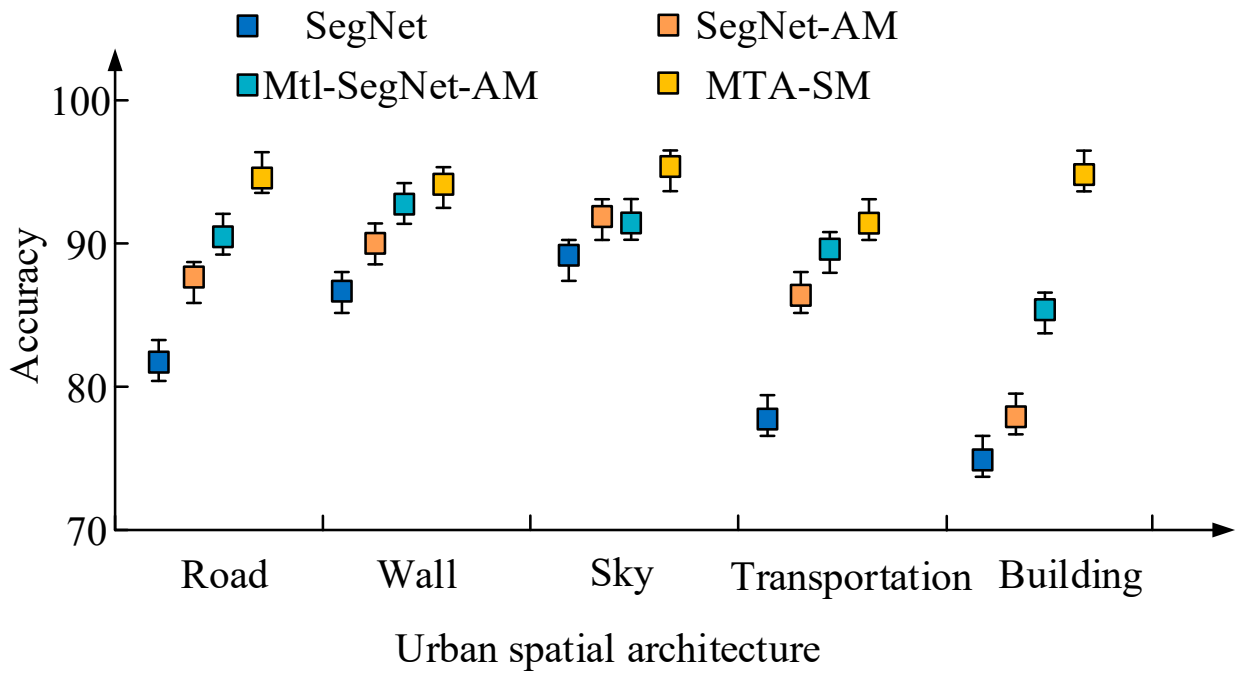

Figure 10. Accurate identification of urban spatial buildings.

As shown in Figure 10, the various algorithms identified walls and sky in cities with high accuracy, and their accuracy in identifying roads and buildings varied. Among them, MTA-SM had accuracies of 95.4, 94.8, 96.2, 92.1, and 96.7 for roads, walls, sky, transportation, and buildings, respectively. The results show that MTA-SM had a higher recognition accuracy for urban spatial buildings compared with the other algorithms. Twenty people living in a city were randomly selected and divided into four groups of five people each. The model recognized the emotions that matched with the five groups of city images. The model's recognition results and its matching performance scores are shown in Figure 11.

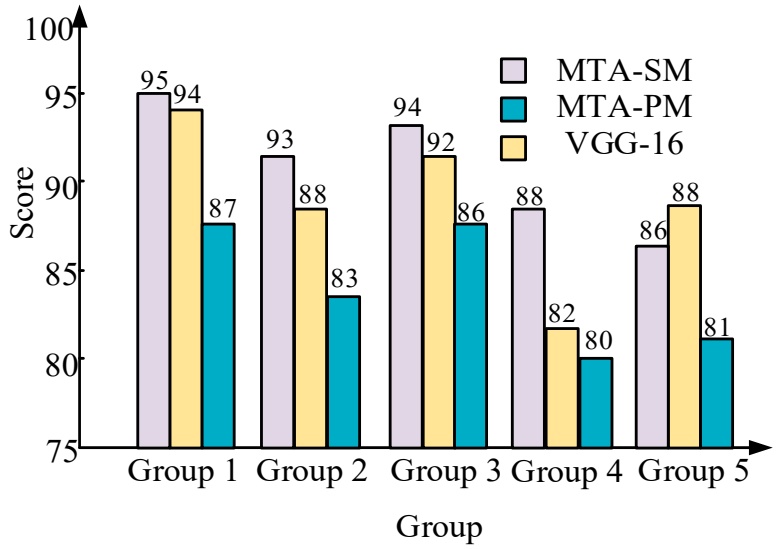

Figure 11. Three methods of user satisfaction survey.

As shown in Figure 11, the first group rated MTA-SM, MTA-PM, and VGG-16 at 95, 94, and 87, respectively. The second group rated MTA-SM, MTA-PM, and VGG-16 at 93, 88, and 83, respectively. The third group rated MTA-SM, MTA-PM, and VGG-16 at 94, 92, and 86, respectively. The fourth group rated MTA-SM, MTA-PM, and VGG-16 at 88, 82, and 80, respectively. The fifth group rated MTA-SM, MTA-PM, and VGG-16 at 86, 88, and 81, respectively. The results show that the ratings of MTA-SM were generally higher than those of the other two methods, indicating that the proposed MTA-SM model had good user satisfaction.

5. Discussion and Conclusions

With the development of machine learning methods for computer vision, computer vision methods that quantify the perception of urban spatial quality can reflect the relationship between urban space and residents' perceptions. This study proposed a multi-task learning urban spatial attribute perception model integrated with an attention mechanism. Using this model, attributes in urban street scenes were identified and matched via semantic segmentation and instance segmentation. The experimental results show that the prediction accuracies of Mtl-VGG were 70.84%, 70.96%, 70.65%, 71.65%, 71.74%, and 71.24%, and the prediction accuracies of MTA-PM were 79.54%, 78.62%, 79.68%, 77.42%, 78.45%, and 76.98%, respectively.

The results of this study show that the proposed MTA-SM model is effective in matching the public's perception of urban spatial quality. This study makes the following contributions: This study provides scientific, data-driven support for creating more accurate and comprehensive guidance for urban planning and design. It can help to create a more livable, sustainable, and human-centered urban environment. Intuitive and objective data are provided, which help us to understand and assess urban spaces more comprehensively.

Nevertheless, this research has several deficiencies. The model constructed in this study requires a large amount of data for training, so it requires a higher computing power. Reducing the computing power requirement and achieving a better model performance would make the proposed model more meaningful for promotion.

Author Contributions: Conceptualization, H.Z.; methodology, C.K.; software, H.L.; validation, H.Z.; formal analysis, H.Z.; investigation, H.L.; resources, C.K.; data curation, H.Z.; writing—original draft preparation, H.Z.; writing—review and editing, C.K.; visualization, H.Z.; supervision, H.L.; project administration, C.K.; funding acquisition, C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant from the Brain Korea 21 Program for Leading Universities and Students (BK21 FOUR) MADEC Marine Design Engineering Education Research Group.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset available from the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, W. Image Recognition Method for Pitching Fingers of Basketball Players Based on Symmetry Algorithm. *Wireless Commun. Mob. Comput.* **2021**, *3*, 2242222. [CrossRef]
2. Gao, P.; Zhao, D.; Chen, X. Multi-dimensional data modelling of video image action recognition and motion capture in deep learning framework. *IET Image Process* **2020**, *14*, 1257–1264. [CrossRef]
3. Wang, R.; Zhao, J.; Meitner, M.J.; Hu, Y.; Xu, X. Characteristics of urban green spaces in relation to aesthetic preference and stress recovery. *Urban For. Urban Green* **2019**, *41*, 6–13. [CrossRef]
4. Wang, D.; Liu, K.; Huang, Y.; Sun, L.; Du, B.; Fu, Y. Automated urban planning aware spatial hierarchies and human instructions. *Knowl. Inf. Syst.* **2023**, *65*, 1337–1364. [CrossRef]
5. Lekus, H.Y. Public Space Humanization in a Night City. *Light Eng.* **2019**, *65*, 28–36. [CrossRef]
6. Hao, Y.; Liang, W.; Yang, L.; He, J.; Wu, J. Methods of image recognition of overhead power line insulators and ice types based on deep weakly-supervised and transfer learning. *IET Gener. Transm. Distrib.* **2022**, *16*, 2140–2153. [CrossRef]
7. Liang, J.; Xu, F.; Yu, S. A multi-scale semantic attention representation for multi-label image recognition with graph networks. *Neurocomputing* **2022**, *491*, 14–23. [CrossRef]
8. Xu, P.; Liu, X.; Zhao, Y.; Lan, D.; Shin, I. Study of graphdiyne biomimetic nanomaterials as fluorescent sensors of ciprofloxacin hydrochloride in water environment. *Desal. Water Treat.* **2023**, *302*, 129–137. [CrossRef]
9. Ma, R.; Li, Z.; Guo, F.; Zhao, L. Hybrid attention mechanism for few-shot relational learning of knowledge graphs. *IET Comput. Vis.* **2021**, *15*, 561–572. [CrossRef]
10. Miller, M.D.; Doherty, J.J.; Butler, N.M.; Coull, W.G. Changing counterproductive beliefs about attention, memory, and multitasking: Impacts of a brief, fully online module. *Appl. Cogn. Psychol.* **2020**, *34*, 710–723. [CrossRef]
11. Xu, P.; Ding, C.; Li, Z.; Yu, R.; Cui, H.; Gao, S. Photocatalytic degradation of air pollutant by modified nano titanium oxide (TiO_2) in a fluidized bed photoreactor: Optimizing and kinetic modeling. *Chemosphere* **2023**, *319*, 137995. [CrossRef] [PubMed]
12. Ahamed, S.A.; Ravi, C. Novel deep learning model for bitcoin price prediction by multiplicative LSTM with attention mechanism and technical indicators. *Int. J. Eng. Syst. Model. Simul.* **2022**, *13*, 164–177. [CrossRef]
13. Tang, J.; Wu, X.; Zhang, M.; Zhang, X.; Jiang, M. Multiway dynamic mask attention networks for natural language inference. *J. Comput. Methods Sci. Eng.* **2020**, *21*, 151–162. [CrossRef]
14. Shi, T.; Qiao, Y.; Zhou, Q. Spatiotemporal evolution and spatial relevance of urban resilience: Evidence from cities of China. *Growth Chang.* **2021**, *52*, 2364–2390. [CrossRef]
15. Cheng, H.H.; Hsu, Y.Y. Integrating spatial multi-criteria evaluation into the potential analysis of culture-led urban development: A case study of Tainan. *Environ. Plan. B Urban Anal. City Sci.* **2022**, *49*, 335–351. [CrossRef]
16. Deng, Y.; Wang, J.; Gao, C.; Li, X.; Wang, Z.; Li, X. Assessing temporal–spatial characteristics of urban travel behaviors from multiday smart-card data. *Phys. A Stat. Mech. Its Appl.* **2021**, *576*, 12–25. [CrossRef]
17. Xu, P.L.; Su, Y.M. Design and implementation of landscape system for East and West Huashi Street in Beijing based on virtual reality technology. *Appl. Mech. Mater.* **2013**, *263*, 1849–1852. [CrossRef]
18. Cao, Y.; Kong, L.; Zhang, L.; Ouyang, Z. Spatial characteristics of ecological degradation and restoration in China from 2000 to 2015 using remote sensing: Ecological degradation and restoration in China. *Restor. Ecol.* **2020**, *28*, 1419–1430. [CrossRef]
19. Lei, Y. Research on micro video character perception and recognition based on target detection technology. *J. Comput. Cogn. Eng.* **2022**, *1*, 83–87. [CrossRef]
20. Suel, E.; Polak, J.W.; Bennett, J.E.; Ezzati, M. Measuring social, environmental and health inequalities using deep learning and street imagery. *Sci. Rep.* **2019**, *9*, 6229–6237. [CrossRef] [PubMed]