International Journal of Geo-Information

MDPI

Article

Extracting the Urban Landscape Features of the Historic District from Street View Images Based on Deep Learning: A Case Study in the Beijing Core Area

Siming Yin, Xian Guo * and Jie Jiang

School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 100044, China; 2108570020075@stu.bucea.edu.cn (S.Y.); jiangjie@bucea.edu.cn (J.J.)
* Correspondence: guoxian@bucea.edu.cn

Abstract: Accurate extraction of urban landscape features in the historic district of China is an essential task for the protection of the cultural and historical heritage. In recent years, deep learning (DL)-based methods have made substantial progress in landscape feature extraction. However, the lack of annotated data and the complex scenarios inside alleyways result in the limited performance of the available DL-based methods when extracting landscape features. To deal with this problem, we built a small yet comprehensive history-core street view (HCSV) dataset and propose a polarized attention-based landscape feature segmentation network (PALESNet) in this article. The polarized self-attention block is employed in PALESNet to discriminate each landscape feature in various situations, whereas the atrous spatial pyramid pooling (ASPP) block is utilized to capture the multi-scale features. As an auxiliary, a transfer learning module was introduced to supplement the knowledge of the network, to overcome the shortage of labeled data and improve its learning capability in the historic districts. Compared to other state-of-the-art methods, our network achieved the highest accuracy in the case study of Beijing Core Area, with an mIoU of 63.7% on the HCSV dataset; and thus could provide sufficient and accurate data for further protection and renewal in Chinese historic districts.

Keywords: street view images; urban landscape; Chinese traditional-style building; deep learning; semantic segmentation; Beijing Core Area

Citation: Yin, S.; Guo, X.; Jiang, J. Extracting the Urban Landscape Features of the Historic District from Street View Images Based on Deep Learning: A Case Study in the Beijing Core Area. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 326. https://doi.org/10.3390/ijgi11060326

Academic Editors: Wolfgang Kainz and Maria Antonia Brovelli

Received: 12 April 2022
Accepted: 27 May 2022
Published: 28 May 2022

1. Introduction

The Beijing Core Area is where Beijing's functions as the nation's political, cultural, and international exchange center are mostly located, as well as being a key area for the preservation of historical districts [1]. The urban landscape (including the landscape of historical heritage and modern life) in this specific area has formed distinct characteristics during the long historical process of its development, which intuitively reflect cultural characteristics, the historical scene, and the esthetic appeal for residents of the city. However, with the impact of urbanization and tourism development, the urban landscape of the Beijing Core Area is undergoing intense changes, and the protection and management of the urban landscape in the core area of Beijing now face serious challenges [2,3].

To meet the demands for urban landscape conservation and sustainable development in the Beijing Core Area, it is necessary to investigate the spatial distribution of typical features of the urban landscape (i.e., the natural features such as the sky, vegetation, and the artificial features like buildings, roads, etc.). In the literature, there are two main kinds of method for landscape feature extraction: field-survey-based methods, and remote-sensing-based methods. A field survey and measurement by manual means usually requires a considerable amount of human and material resources [4–7]. Remote-sensing-based methods are efficient in obtaining the spatial distribution of the landscape over a large area [8,9]. However, conventional aerial or satellite images cannot acquire the side and

façade information of the landscape features [10], while the recently booming technology of oblique photogrammetry [11] faces a data shortage problem in Beijing Core Area, due to the established no-fly zone. With the popularization of location-based services, street view images are accessible and, in recent years, have been attracting increasing attention with regard to autonomous driving [12], urban environment studies [13], and urban landscape investigations [14,15]. Compared with aerial photographs and satellite images, street view images are "people-oriented" and provide the potential to obtain more information [16], making them suitable for extracting detailed landscape features in complex areas of the Beijing Core Area.

Nevertheless, the rich and diverse detailed information on the urban landscape embedded in the street view images generates higher requirements for landscape feature extraction methods. Traditional image-analysis-based methods, such as pixel-based methods [15], object-based methods [17], and scene-based methods [18], used hand-crafted features to characterize urban landscape features, but failed to extract high-level semantic features from street view data. With recent advances in deep neural networks (DNNs), various DNN-based models have been proposed for semantic segmentation [19–25], which can automatically derive features that are tailored to the segmentation tasks [26], which makes such methods better choices for handling complicated scenarios, especially in street view images [27,28].

Thanks to the effectiveness of semantic segmentation, several researchers have attempted to extract landscape features in street view images. Gong et al. [12,13] extracted three typical landscape features (i.e., vegetation, building, and sky) from Google Street View data using PSPNet [29], and used them to calculate the tree view factor (TVF), building view factor (BVF), and sky view factor (SVF) in the downtown of Hongkong. Middel et al. [30] extracted six landscape features from Google Street View data to derive street-level morphology and urban feature composition, as experienced by a pedestrian. Ye et al. [31] extracted multiple urban landscape features using SegNet [21] from Baidu Street View data of central Shanghai, to estimate the visual quality of urban streets. Suel et al. [32] proposed a novel deep learning based multimodal framework to jointly utilize satellite and street-level images for measuring income, overcrowding, and environmental deprivation in urban areas. Recently, Zhang et al. [33] proposed TBMask R-CNN, to extract Chinese traditional-style buildings within the Fifth Ring Road of Beijing from Tencent Street View data and quantify pedestrians' visual perceptions of the traditional buildings. However, information on the historic districts of China is rare or nonexistent in previous datasets, making it impossible to extract historical landscape features. In addition, the existing semantic-segmentation-based methods suffer from the high variety of semantically complex contents in the historic districts, due to the high flow of people and traffic, narrow alleyways, numerous details, and the changeable lighting conditions.

In the most recent studies, an attention mechanism was developed to improve the performance of the models in complex scenarios, by enhancing the important part of the input data and fading out the rest [34]. Inspired by the classical non-local means, an asymmetric non-local neural network (ANNN) was proposed to improve the image recognition result [35]. Zhao et al. [22] proposed a point-wise spatial attention network (PSANet) to harvest contextual information from all positions in the feature maps, by connecting each position with all others through a self-adaptively learned attention map. Although attention-based models have achieved state-of-the-art performance in complex scenarios, they usually require a large amount of labeled data for model training, even more than a conventional semantic segmentation network. Furthermore, the implementation of the attention mechanism also caused unstable output, and thus requires more iterations in the training phase. Thus, attention-based methods have not been utilized to extract landscape features as far as we know. In addition, there is no existing dataset developed specifically for landscape feature extraction in the center of Beijing, and the existing datasets (e.g., Cityscapes) cannot effectively supply detailed information on the urban landscape with Chinese characteristics.

To address these challenges, in this paper, we built a small yet comprehensive history-core street view (HCSV) dataset, which is composed of fine labels for every typical land feature in the core area. Furthermore, we introduced transfer learning technology, polarized self-attention (PSA) block, and atrous spatial pyramid pooling (ASPP) block to enhance the performance of the DNN models, regarding the complex environment with relatively limited annotations. Finally, we evaluated our method on the HCSV dataset and compared it with multiple state-of-the-art segmentation networks. As far as we know, this is the first time that an automatic landscape feature extraction method has been developed specifically for the historic districts in China, such as the Beijing Core Area. We summarize our main contributions as follows:

- We construct a novel dataset for the Beijing Core Area;
- We introduce transform learning techniques and a PSA attention block to improve the performance of the network in complex environments and small-sample scenarios; and
- We verify the proposed method on the HCSV dataset and compare it with other state-of-the-art deep learning methods in the Beijing Core Area.

The paper is organized as follows: Section 2 describes the study area and proposed dataset, followed by the methodology being presented in Section 3. Section 4 conducts a detailed experiment. Section 5 discusses the proposed dataset and explores the best segmentation architecture for the Beijing Core Area. Finally, we conclude this article in Section 6.

2. Historical-Core Street View (HCSV) Dataset

In this section, we introduce our proposed Historical-Core Street View dataset by clarifying the study area and the procedure of data annotation.

2.1. Study Area

As the capital of five imperial dynasties (Liao, Jin, Yuan, Ming, and Qing) and the current capital, Beijing, nestled on the north of the North China Plain, covers an area of 16,410.54 square kilometers (sq. km). The traditional residential area of Beijing has been recognized as a world cultural heritage site with unparalleled historical, cultural, and social value. The area with a high concentration of historic buildings and alleyways in the traditional residential zones in Beijing, namely the Beijing Core Area, is regarded as a conveyer of Chinese history and culture, and is considered a typical representative of a Chinese historic district. Tremendous efforts have been devoted to the protection and management of this area by the local government.

During the study on the Beijing Core Area, we discovered that the whole area could be divided into three categories: modern street, modern residential, and ancient alleyway, respectively. As one of the oldest residential blocks and the most famous attraction in the Beijing Core Area, South Luogu Lane contains multiple historical alleyways (e.g., Maoer Hutong) and is surrounded by modern avenues (e.g., Di'anmenwai Street). Therefore, Di'anmenwai Street and Mouer Hutong were selected as the study areas to cover the main scenery shown in the core area. The details are shown in Figure 1 below.

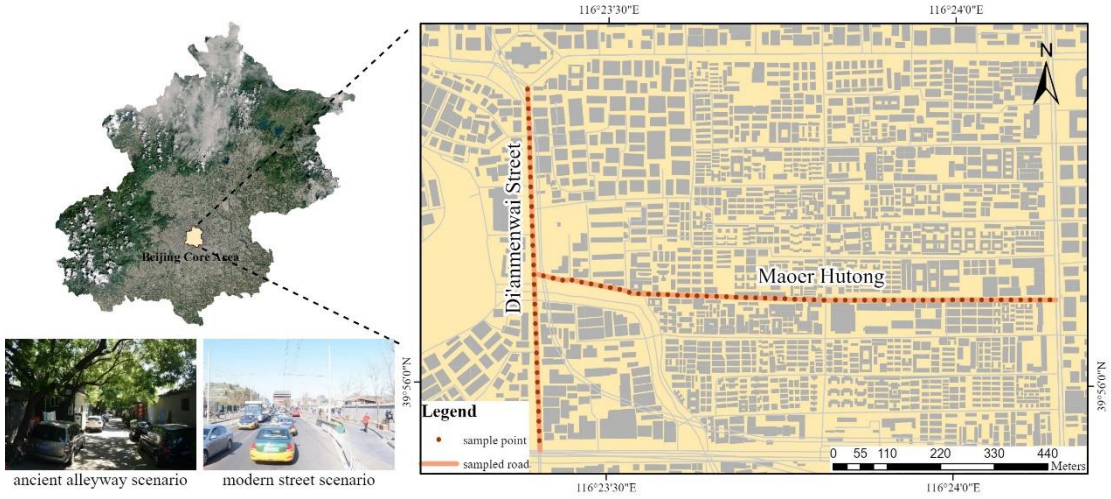

Figure 1. Study area.

2.2. Dataset

With the development of real-world map services in China, multiple local Internet mapping providers such as Baidu and Tencent have made street view images widely available, which represent street façade information from the pedestrian perspective, and thus provide a new data source for the study of the urban landscape. Based on street view data, various datasets have been developed for different purposes, and we summarize them in Table 1 with respect to image volume, categories, and locations.

Table 1. Information of different datasets related to semantic segmentation.

Dataset	Year	Categories	Images	Location
CamVid	2007	32	701	Britain urban and residential
ImageNet (pretrain)	2009	20 k	14 m	Unknown
Cityscapes	2016	30	5 k (fine)	Europe
ADE20k	2017	150	20 k	Unknown
COCO (pan)	2018	171	118 k/5 k (train/val)	Unknown
ApolloScapes	2018	24	140 k	China
Bdd100k	2020	40	10 k	United States

Evidently, most of these datasets have thousands of annotated images to train modern deep learning models. As the largest visual dataset in the world, the ImageNet [36] dataset has more than 14 million images, all comprehensively annotated in 20 thousand categories. Although it cannot directly train for a down-stream task such as semantic segmentation, it has been widely applied for the pretraining procedure, due to the vast amount of common visual information. The CamVid dataset [37], which has 701 dense-labeled images acquired from video sequences, was one of the most commonly used road scene understanding datasets for the early study in semantic segmentation. The Cityscapes dataset [38] comprises 5000 finely annotated street view images of 50 cities in Europe from selected driving video frames. To further enrich the scenarios for scene understanding, the ADE20k dataset contains more than 20,000 annotated images with 150 different categories. The COCO dataset, on the other hand, contains a total of 123 thousand images and released the annotation for panoptic segmentation in 2018, which has 171 different categories [39]. However, this dataset includes multiple indoor and outdoor scenes and does not concentrate on street scenarios. By contrast, a few datasets were announced specifically for street scenarios, including the Bdd100k dataset [40] and the ApolloScapes dataset [41].

Although these datasets have provided plenty of annotated images in a street scenario, there are still improvements that could be made for the sake of the urban landscape features extraction task. First, most existing datasets acquired street view data by recording forward videos along a street, which could benefit tasks such as autonomous driving but cannot satisfy landscape extraction needs, due to the lack of multi-view (or panorama) information. Apart from that, none of the existing datasets contain unique landscape feature information (for example, traditional buildings and variable details, etc.) for the Chinese historic district. Thus, we first collected the relevant streetscape images, and then set up a semantic segmentation annotation platform and constructed the first street view dataset for the core area in Beijing.

The developed HCSV dataset complements the existing dataset in terms of offering different views of street view images and annotating them with a specific class set about landscape features in the Beijing Core Area. This dataset contains 127 pixel-level annotated images taken from Di'anmenwai Street and Maoer Hutong, which contain the three typical scenarios (i.e., modern street, modern residential, and ancient alleyway) of the urban landscape in the Beijing Core Area.

The images in the HCSV dataset were obtained from Baidu Map Service, which supplies an application programming interface (API) for querying and downloading street view images with multiple parameters, e.g., size, coordinate, heading, pitch, and field of view (FOV). Specifically, we collected the street view images (with resolution of 512×1024) in four directions (with a pitch angle of 0° and headings of 0°, 90°, 180°, and 270°, with the FOV set to 90°) for each sample point along the streets or alleyways at 20-m intervals. After data collection, a data cleaning process was applied using the mean hash method, to detect and remove all duplicated images.

For annotation, we developed an online semantic segmentation editor platform (shown in Figure 2) based on an open source project (https://github.com/Hitachi-Automotive-And-Industry-Lab/semantic-segmentation-editor, accessed on 18 May 2022) [42]. Furthermore, we established a custom class set to summarize the distinctive urban landscape features, especially for Chinese historic districts. The class set was simplified (integrating the objects with similar semantics, e.g., sign and traffic light) and extended (expanding the semantics, e.g., archways and other constructions were included in the "building" class and we set up novel classes such as clutter) from the original Cityscapes class set. Thus, our dataset could outline the landscape for every typical scenario inside the Beijing Core Area, while keeping a simple category hierarchy, compared to existing datasets. The categories in our dataset are shown in Table 2 and we show the typical landscape features for each category in Figure 3. This HCSV dataset will be made openly available for all research needs.

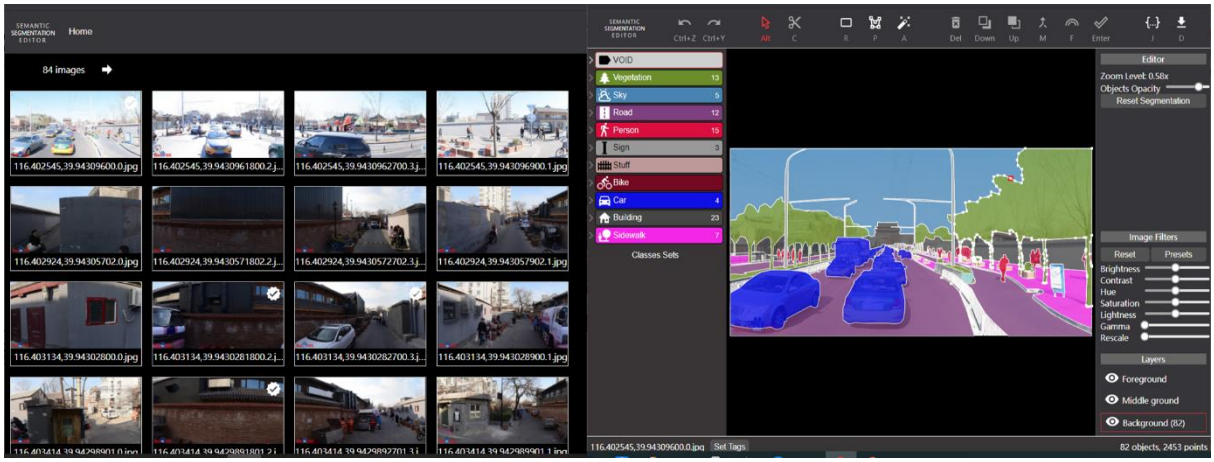

Figure 2. Semantic segmentation editor platform.

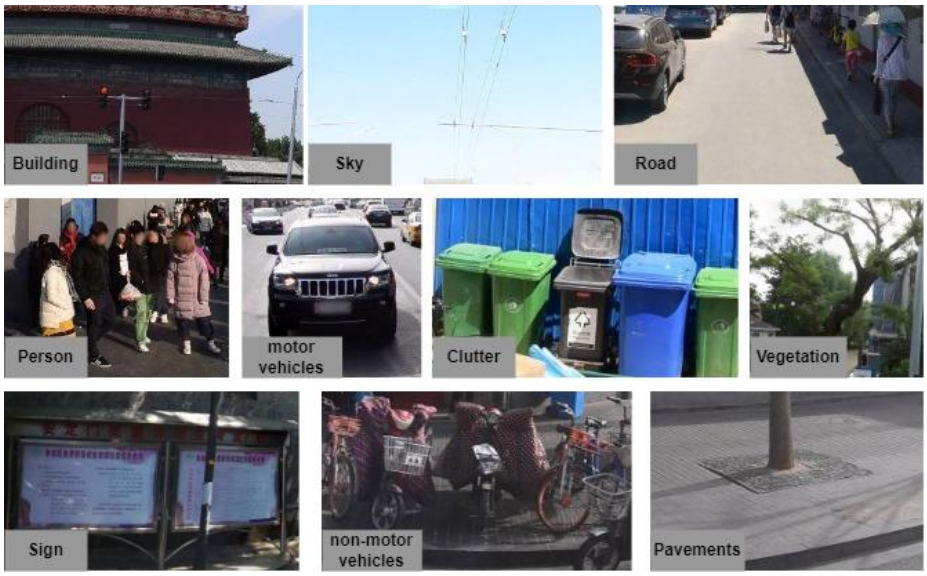

Figure 3. The typical urban landscape features for each category in the HCSV dataset.

Table 2. The class set of the HCSV dataset.

Class Name	Description
vegetation	The vegetation along the street
sky	The unobstructed sky in street view images
road	The surface of the main road or alleyways
person	People and what they are holding
sign	Street signs, streetlights, and traffic signals
clutter	Trash, abandoned furniture, and tricycles
non-motor vehicles	Bicycles
motor vehicles	Cars, busses, etc.
building	Traditional courtyards, modern buildings, and other constructions
pavements	Sideways along the main road

3. Methodology

The architecture of the proposed method, i.e., PALESNet, is shown in Figure 4. Aiming at extracting urban landscape features of the Beijing Core Area, particularly with a limited number of samples, the proposed network consists of three parts: a feature extractor module, a pixel-level segmentation module, and a knowledge transfer module. To distinguish different kinds of landscape features in the Beijing Core Area, the feature extractor first extracts features from the inputted street view images. Then, the pixel-level segmentation module learns the multi-scale semantic representation from the features previously extracted and segments each landscape feature. In this module, the PSA block is utilized to discriminate landscape features more effectively, especially in a complex environment, whereas the ASPP block is utilized to capture the multi-scale features. In addition, a knowledge transfer module is applied to enhance the recognition performance under a limited sample situation.

Let I represent the street view image from HCSV, and let y be the label of the landscape features. The flowchart of this study can be summed up as follows:

1. First, the image I is input into the feature extractor module, to obtain a group of feature vectors $F = \{feat^1, feat^2, feat^3, feat^4\}$.
2. Next, the feature vector F is taken into the pixel-level segmentation module to identify each class of urban landscape features, where the PSA block and ASPP block are implemented to make F more discriminative and extract the multi-scale features, respectively. The cross-entropy loss (CEL) is then calculated according to the output

of the network R and the label y. Moreover, a data augmentation algorithm is utilized to increase the training sample and enhance the robustness of the model.

3. In the training phase, the knowledge transfer module is activated to initialize the parameters of the network $p = \{p_e, p_s\}$, where p_e is the parameter of the feature extractor and p_s stands for the parameter of the segmentation module. The initialized parameter p comes from the well-pretrained model with knowledge of the existing dataset [36,38].

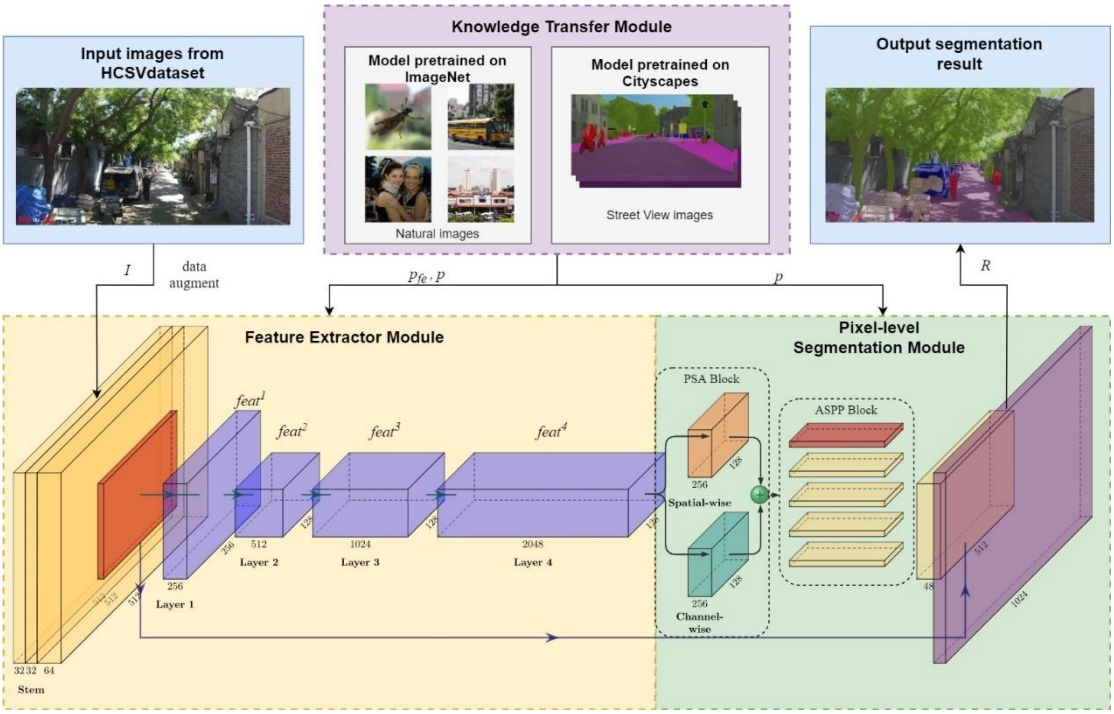

Figure 4. Flowchart of the proposed PALESNet. A feature extractor module is used to extract features from the street view images. Features are sent to the segmentation module to recognize landscape features, where the PSA block is utilized to make the feature more discriminative and the ASPP block is used to capture the multiscale representation of each feature. The knowledge transfer module is activated during the training phase, to initialize the parameters for both the feature extractor and the segmentation network.

3.1. Feature Extractor

To identify each landscape feature in the Beijing Core Area, we needed to extract features from street view images. Many previous studies have proven that deep CNN (DCNN) has a strong ability for feature extraction, and thus we built our feature extractor based on ResNet-r50 [43]. In addition, a data augmentation procedure was introduced to yield more training samples before the street view images I are input into the network, which includes the operation of random crop and resize, random rotation, and random distortion. Apart from the increase in the sample volume, the data augment procedure can enhance the robustness of the network, by providing random-operated samples.

To increase the depth of the traditional CNNs, while overcoming problems of gradient disappearance and gradient explosion, the ResNet is composed by the architecture called residual block (Figure 5), which utilizes a shortcut connection to transfer the input x directly to the output. The output of the residual block is as follows:

$$F(x) + x = \sigma(W_3(\sigma(W_2(\sigma(W_1(x))))) + x) \quad (1)$$

where W_1, W_2, and W_3 represent the weight of the convolution layers, respectively, whereas σ is the rectified linear unit (ReLU) function [44]. The shortcut connection could perform

identity mapping more effectively than the plain CNNs, and thus can resolve the degradation problem which limited the depth of the network. Furthermore, the residual architecture can avoid the phenomenon of gradient vanishing, by carrying the gradient throughout the extent of the DCNN [45].

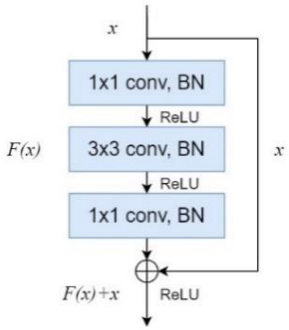

Figure 5. The structure of the residual block.

After removing the average pooling layer and the fully connected layer from the original ResNet-r50, a variant of ResNet for feature extraction was obtained. Our feature extractor consists of five stages. For the first stem stage, the original 7×7 convolution layer of ResNet is replaced by three 3×3 convolution layers, as shown in Figure 6. The reason for the replacement was to reduce the computation cost of the large convolution kernel. For example, the 7×7 kernel with k filters is 5.4 times more expensive than the 3×3 kernel with the same number of filters. To address the lower expressiveness caused by the reduced kernel size, three 3×3 convolution layers are stacked serially to extract more features. In our case, the modified stem stage only has $(2 \times 3 \times 3 \times 32 + 3 \times 3 \times 64)/(7 \times 7 \times 64) = 0.37$ times the computation of the original stem stage.

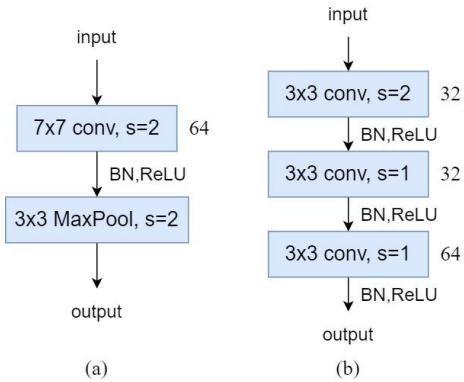

Figure 6. The stem stage of the (**a**) original ResNet and (**b**) our feature extractor.

Following the stem stage are four residual layer stages, which are composed of 3, 4, 6, and 3 residual blocks, respectively. Note that each block is gradually deepened and reaches the depth of 2048 at the end. The output feature $feat^4$ of the feature extractor is 1/8 of the original input image size.

3.2. Pixel-Level Segmentation Module

The segmentation module was designed to recognize each landscape feature at a variable scale (i.e., the scale between different features, and scale variability caused by different distances), which is quite common in Chinese historic districts. Furthermore, the segmentation module is required to distinguish different features under the complex environment, which is caused by uneven lighting conditions, high traffic flow, and narrow alleyways. Therefore, we introduced the PSA to discriminate features in the complex

environment and utilize the atrous spatial pyramid pool (ASPP) to fuse the multi-scale features.

3.2.1. Polarized Self Attention

The PSA block can capture the long-range dependency and make features more discriminative on the feature maps, and is a very lightweight model that does not require excessive costs regarding memory and calculation, while keeping a high resolution in attention computation. As Figure 7 shows, the PSA contains two submodules: the channel-only self-attention module, and the spatial-only self-attention module.

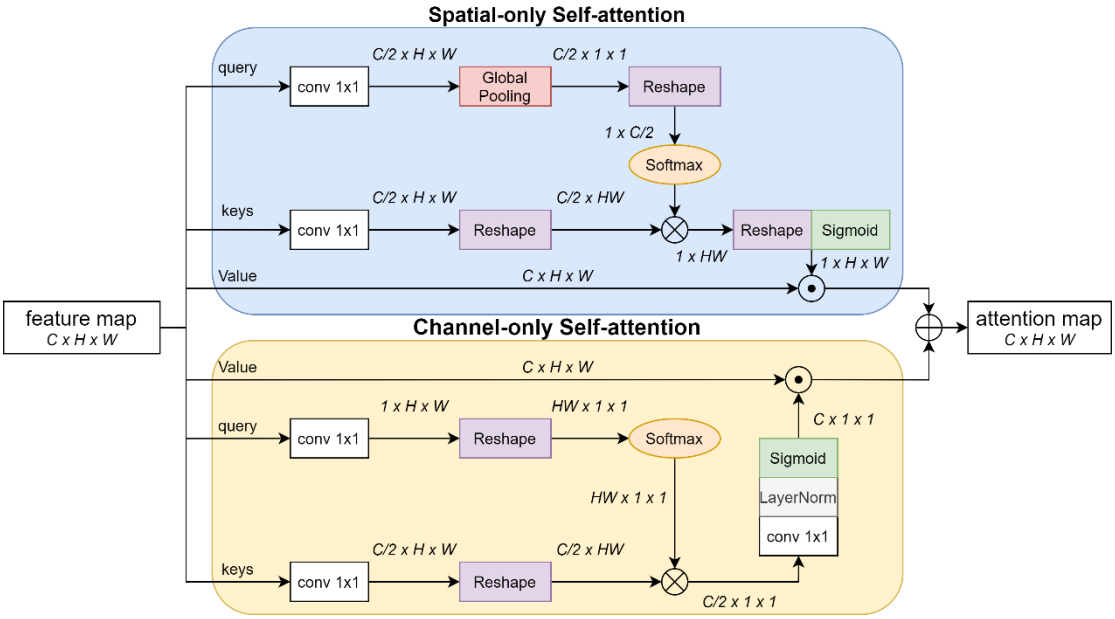

Figure 7. The architecture of the polarized self-attention (PSA) module. The input feature is sent to the spatial-wise attention and channel-wise attention parallelly, to make it more discriminative for a complex environment.

The channel-only self-attention module aims to capture the long-range dependency through a channel attention map and highlight the class-specific features, which can be calculated using the following formula:

$$P^{ch}(X) = f_{SG}[W_c(\varphi_1(W_a(X)) \times \sigma(\varphi_2(W_b(X))))] \tag{2}$$

where X denotes an input feature of size $C \times H \times W$, and f_{SG} represents the sigmoid function. The W_a, W_b, and W_c are 1×1 convolution layers, respectively, φ_1, φ_2 are two tensor reshape operators, and σ is the SoftMax function. The $P^{ch}(X)$ is the channel attention map and "$\times$" is the matrix dot-product operation.

Similarly, the spatial-only self-attention module also applies 1×1 convolution at first and then reshapes the result. Unlike the channel-only branch, the global pooling function f_{GP} is adopted after the first convolution, to compress the redundant information, followed by the SoftMax function σ. Then, the matrix dot-product operation is applied between two internal tensors. Finally, a sigmoid function f_{SG} is used to obtain the final spatial attention map $P^{sp}(X)$. The formula can be described as follows:

$$P^{sp}(X) = f_{SG}[\varphi_3(\sigma(\varphi_1(f_{GP}(W_b(X)))) \times \varphi_2(W_a(X)))] \tag{3}$$

A further highlighted feature F will be generated after the channel-only self-attention module and the spatial-only self-attention module. In this article, we selected the parallel structure of PSA, which can be described as:

$$F(X) = P^{ch}(X) \odot^{ch} X + P^{sp}(X) \odot^{sp} X \tag{4}$$

where $\odot^{ch}$ and $\odot^{sp}$ are the channel-wise and spatial-wise multiplication operators, respectively, and "+" denotes the element-wise addition operator. In our case, the output of the feature extractor, i.e., $feat^4$, was put through the PSA block to obtain a more discriminative feature $feat' = F(feat^4)$ both spatial-wise and channel-wise, which could benefit the subsequent decoding in the complex scenario in the Beijing Core Area.

3.2.2. ASPP Block

As discussed above, the scale of the landscape features is variable in the Beijing Core Area. Thus, we introduced the ASPP block into our network, which has shown promising results on multiple semantic segmentation models [23,29]. As shown in Figure 8, the ASPP block can extract multi-scale features generated by the feature extractor and computed by the previous PSA block.

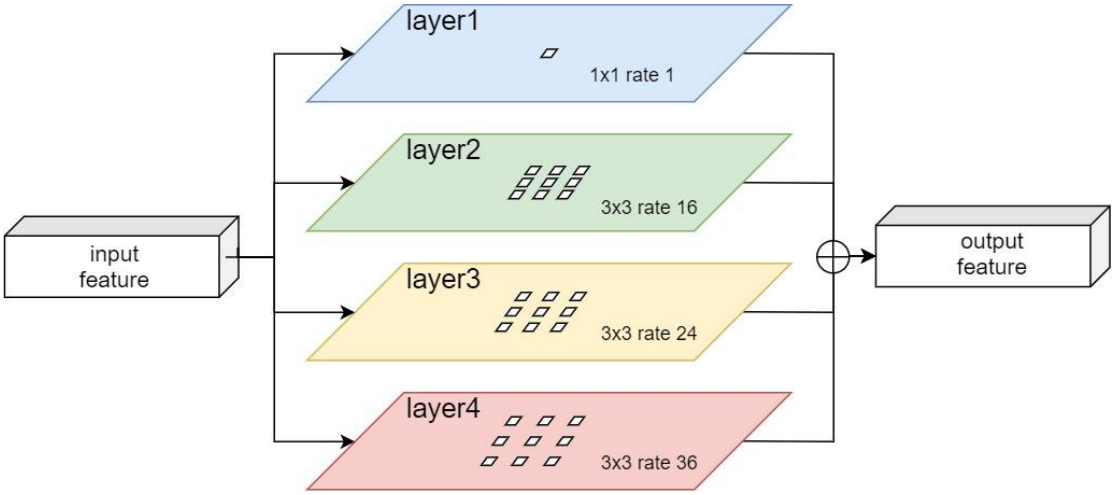

Figure 8. The illustration of the ASPP block, in which '⊕' denotes the concatenate operation.

The ASPP block consists of four different settings of the receptive field. The first layer is composed of a convolution layer with the kernel size of 1×1, followed by a batch normal layer and a ReLU activation layer. For the rest of the layers, the kernel size of the convolution layer was set to 3×3, with a dilation rate of 16, 24, and 36, respectively. The atrous convolution with different dilation rates results in the different receptive fields, which capture multiscale features effectively and efficiently, because it requires fewer parameters than the ordinary convolution operation (e.g., convolution with a kernel size of 16×16) to achieve a larger receptive field. The previous feature $feat'$ is further exploited through the aforementioned layers parallelly and concatenated to form the output feature $feat^*$. Finally, $feat^*$ is quadruple upsampled using the bilinear sampling method.

Finally, the feature $feat^*$ is decoded to obtain the semantic information of each landscape feature in the street view image I at the pixel level. Inspired by the latest work of DeepLab series, the low-level features $feat^1$ will concatenate with $feat^*$ after a 1×1 convolution bottleneck, to recover the spatial detail in street view images, and then bilinearly upsampled by a factor of 4. Through the combination of the 3×3 and 1×1 convolution layers, the final pixel-level segmentation map of landscape features is obtained.

3.3. Transfer Learning Module

Extracting landscape features accurately in complex scenarios in the Beijing Core Area, especially with the limited number of labeled samples, remains a challenging problem for most existing semantic segmentation methods. To solve this problem, we introduced transfer learning technology, which can transfer knowledge from the relevant existing datasets to our proposed network.

As the largest visual dataset in the world, ImageNet has immense quantities of images about common items, while Cityscapes (mentioned in Section 2.2) shares lots of common features in modern urban scenes, and thus we could employ them both to enrich the knowledge of the proposed network. Thus, we introduced a two-step transfer learning strategy in this paper.

3.3.1. Transfer of Knowledge from ImageNet

The feature extractor is used to extract low-level features embedded in images, such as a boundary feature, which always share similar characteristics in different fields of images (i.e., nature images and street view images). Therefore, it is possible to transfer common knowledge across heterogeneous image domains.

In our case, the transfer of knowledge means transferring the parameters from one well-trained network to the target network, which has similar but different data domains and tasks. Let D_{img} represent the domain of common scenes of ImageNet and T_f be the feature extraction task for the ordinary items, and let D_{hcsv}, $T_f{}'$ represent the domain in HCSV and expected feature extraction task for landscape features, respectively. The process of transfer can be described as follows:

$$P_{fe}\left\{D_{img}, T_f\right\} \xrightarrow{transfer_1} P_{fe}{}'\left\{D_{hcsv}, T_f{}'\right\} \tag{5}$$

where P_{fe} denotes the parameters of the feature extraction model trained by ImageNet for feature extraction purposes, and $P_{fe}{}'$ is the parameters of our feature extractor. After this transfer progress, the parameters of the proposed feature extractor are initialized by the pretrained model (with the same architecture) in ImageNet, and thus receive the knowledge of ordinary items in natural images before being actually trained on the HCSV dataset. Such knowledge could strengthen the performance of the feature extractor in a limited sample situation.

3.3.2. Transfer of Knowledge from Cityscapes

To enrich the knowledge of the modern urban scenes for the proposed network, we transferred the knowledge embedded in the Cityscapes dataset. The procedure of knowledge transfer can be described as follows:

$$P\left(P_{fe}, Seg_{cs}\right) \xrightarrow{transfer_2} P'\left(P_{fe}{}', Seg_{hcsv}\right) \tag{6}$$

where Seg_{cs} and Seg_{hcsv} denote the segmentation task for the objects in Cityscapes and landscape features in HCSV, respectively. P and P' denote the parameters of the proposed network for the task of Seg_{cs} and Seg_{hcsv}.

Once the network had gained the information about the general feature of the ordinary items, as well the characteristics in metropolis streets, the proposed HCSV dataset was then applied to train the networks, to learn the specific feature in the core area of Beijing. It is worth noting that layer1, layer2, and layer3 of the feature extractor are frozen after the $transfer_2$, to reserve the ability to capture the common features, as well as accelerate the fitting process, especially in this limited sample situation.

4. Experimental Results

In this section, the dataset and comparison algorithms employed in the following experiments are first illustrated. Then, the implementation details and evaluation metrics are briefly provided. Finally, the results of the experiments are analyzed in detail.

4.1. Dataset

In this experiment, the proposed HCSV dataset was used to evaluate the performance of our method and make a comparison between it and other CNN-based segmentation models. As discussed in Section 2, the HCSV dataset contains 127 manually annotated street view images in the Beijing Core Area, with a size of 512 × 1024. These images were fully annotated and selected carefully to cover the environmental scenarios (e.g., wide streets, narrow alleyways, crowded residential areas, etc.) presented in the study area. In our experiments, the ratio of samples utilized for the training, validation, and test sets was set to 8:1:1.

4.2. Comparative Methods

To verify the validity of our method, three state-of-the-art segmentation models were chosen for comparison purposes and introduced in brief:

1 Fully convolutional network (FCN): FCN [19] replaces the fully connected layers with a 1 × 1 convolution layer at the end of the general CNN architecture. We introduced this classical network as the baseline that did not contain any extra modification (i.e., attention mechanism, ASPP block, etc.)

2 Asymmetric non-local neural network (ANNN): Inspired by the classical non-local means, ANNN [35] was developed as a simple feedforward block for computing non-local filtering, which can directly capture long-range dependencies while maintaining the variable input sizes and can be easily combined with other operations. ANNN-based building blocks have been applied to efficiency-focused computer vision architectures (such as video classification and segmentation).

3 Point-wise spatial attention network (PSANet): With a similar purpose of capturing long-range context dependencies as ANNN, PSANet [22] uses a novel point-wise bi-direction information aggregation block to capture the contextual information, and insert this block into the conventional FCN. This network had achieved the top performance at the time on various datasets, including Cityscapes and ADE20K, demonstrating its effectiveness and generality, and thus it was chosen to represent the state-of-the-art attention-integrated networks.

In summary, a classical network (FCN) was utilized as the baseline segmentation network, while two models (ANNN and PSANet) were employed as the CNN-based attention-integrated networks, which could be beneficial for landscape feature recognition.

4.3. Experimental Setup and Evaluation Metrics

Experiments were conducted using PyTorch 1.6.0 with the Python 3.7 library on a machine equipped with an Intel Xeon E3-1200 (QuadCore), 32-GB RAM, and an Nvidia GeForce GTX Titan X (12-GB RAM). In the experiments, all the models were built upon a unified benchmark platform, MMSegmentation, which provided a modular design to construct a customized semantic segmentation framework, while supporting multiple contemporary semantic segmentation frameworks for fair competition [46]. We applied the same feature extractor (i.e., ResNet-r50 backbone) for all comparison models. The parameters used in the ImageNet transfer module were officially provided by the PyTorch. The iteration of the training process in the Cityscapes transfer module was set to 40,000 in the Cityscapes dataset, and then to 20,000 for training with the HCSV dataset. We optimized the parameters for all comparison models using the SGD optimizer with a momentum of 0.9 and a poly learning rate policy that decayed from 0.01 to 0.0001, where the weight decay was set to 0.0005.

To evaluate the performance of the comparative methods, four metrics were used: the overall accuracy (*aAcc*), intersection over union (*IoU*) for each class, the mean intersection over union (*mIoU*), and the mean accuracy (*mAcc*), to evaluate the segmentation accuracy. These indices were calculated as follows:

$$aAcc = \frac{\sum_{i=1}^{n} X_{ii}}{M} \tag{7}$$

$$IoU = \frac{X_{ii}}{\sum_{j=1}^{n} X_{ij} + \sum_{j=1}^{n} X_{ji} - X_{ii}} \tag{8}$$

$$mIoU = \frac{1}{n} \sum_{i=1}^{n} \frac{X_{ii}}{\sum_{j=1}^{n} X_{ij} + \sum_{j=1}^{n} X_{ji} - X_{ii}} \tag{9}$$

$$mAcc = \frac{1}{n} \sum_{i=1}^{n} \frac{X_{ii}}{N_i} \tag{10}$$

where X_{ij} denotes the number of pixel class i predicted as class j. Let n be the number of classes and M be the total number of pixels, while N_i represents the total pixels of the designated class i.

For the efficiency assessment, the theoretical and the practical influences are considered. The number of the model parameter and the floating-point operations (*FLOPs*) is utilized as the theoretical indices. The memory and computation power required by each model could be evaluated using these indices individually. Among them, the *FLOPs* are calculated as follows:

$$FLOPs = 2HW\left(C_{in}K^2 + 1\right)C_{out} \tag{11}$$

where H and W are the height and width of the input feature maps, and the K is the size of the kernel size of the convolution process. C_{in} and C_{out} denote the channel of the input or the output feature maps, respectively. As for the practical efficiency evaluation, we recorded the GPU memory occupation and the speed of the prediction process.

4.4. Results on the HCSV Dataset

For a fair comparison with other semantic segmentation models under the same prior knowledge, we transferred the same knowledge for all methods with the transfer learning module before testing, and therefore the main difference lay in the architecture of the networks. Table 3 presents the quantitative results for comparison metrics of all methods. We can see that the classic FCN without attention achieved a comparative performance with ANNN, which integrates a non-local attention block, and outperformed the PSANet by a large margin. This indicates that the different strategies of attention will have a large impact on the result and that the previous attention strategy is not suitable for the landscape features extraction task in the Beijing Core Area. The possible reason for the unsatisfactory performance of ANNN and PSANet is that they need relatively more information for the segmentation in a complex environment such as the Beijing Core Area, which would be even worse in a small-sample situation. Our method, which aggregates multi-scale information and discriminative features by utilizing the PSA module and ASPP module, achieved the best performance in terms of the three evaluation metrics (63.70%, 72.77%, 90.47%), and thus is more suitable for extracting variable features in complex scenarios.

Table 3. The average performance of the different segmentation CNNs on the HCSV dataset.

Model	*mIoU*	*mAcc*	*aAcc*
FCN	59.32	69.63	88.44
PSANet	54.59	64.7	86.78
ANNN	59.1	68.4	88.3
PALESNet (ours)	63.7	72.77	90.47

Although there are various landscape features in the Beijing Core Area, these can be summarized into three typical scenarios: modern street, modern residential, and ancient alleyway scenario. The segmentation results for each comparison method in every typical scenario is shown in Figure 9, and we give the category-by-category segmentation results in Figure 10. As we can see, in the modern residential scenario, all models achieved reasonable results, which indicates the effectiveness of the transfer learning procedure that successfully transferred the knowledge of urban landscape from ImageNet and Cityscapes. In this specific scenario, there are many different types (the flowerpot on the road, tricycle, etc.) of the class "clutter", which led to an unstable segmentation result. It can be seen that only our method distinguished the flowerpot from the crowd, which indicates that our method can learn new feature representations more effectively than the others. The possible reason is that the frozen strategy in the feature extractor limited the parameter space and made the network converge more easily during the training phase.

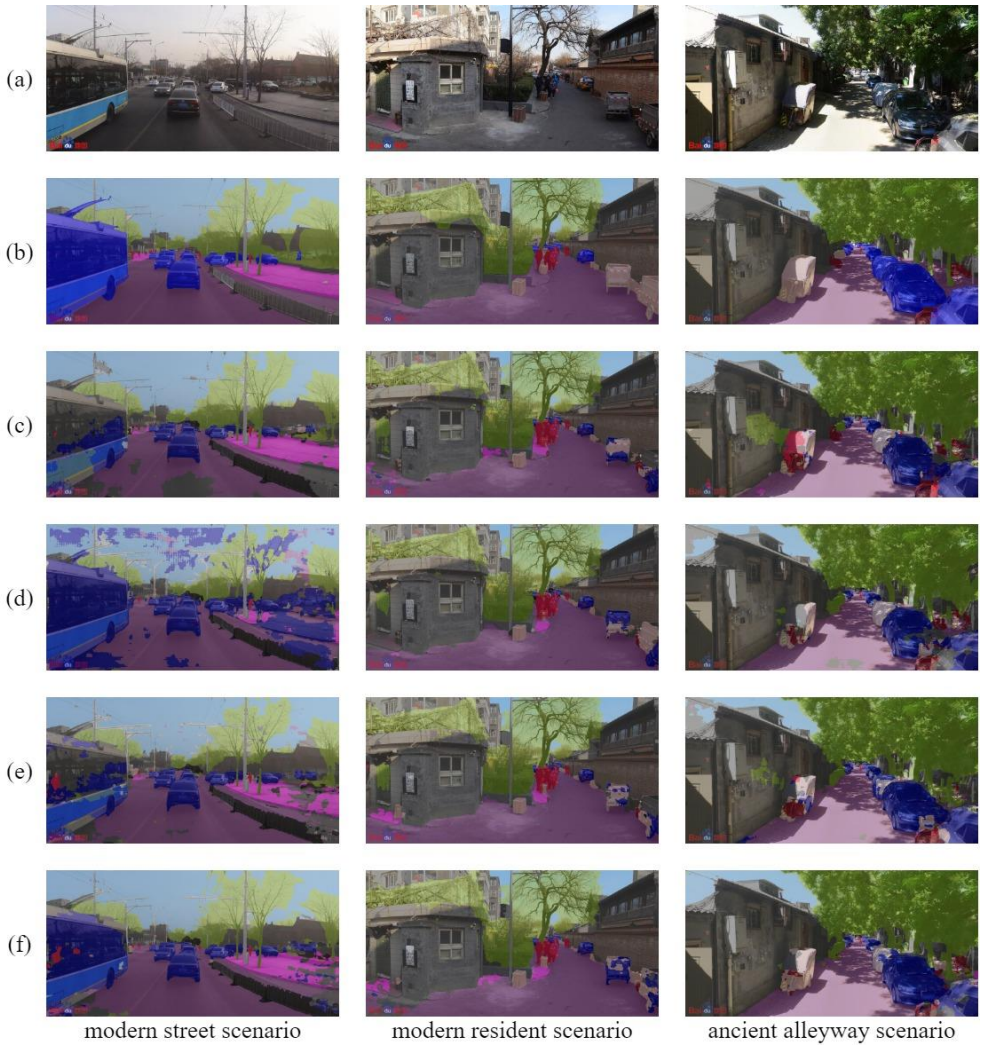

Figure 9. Visual comparison of the modern street scenario, modern residential scenario, and ancient alleyway scenario on HCSV dataset: (**a**) original image; (**b**) ground truth; (**c**) FCN; (**d**) PSANet; (**e**) ANNN; (**f**) our method.

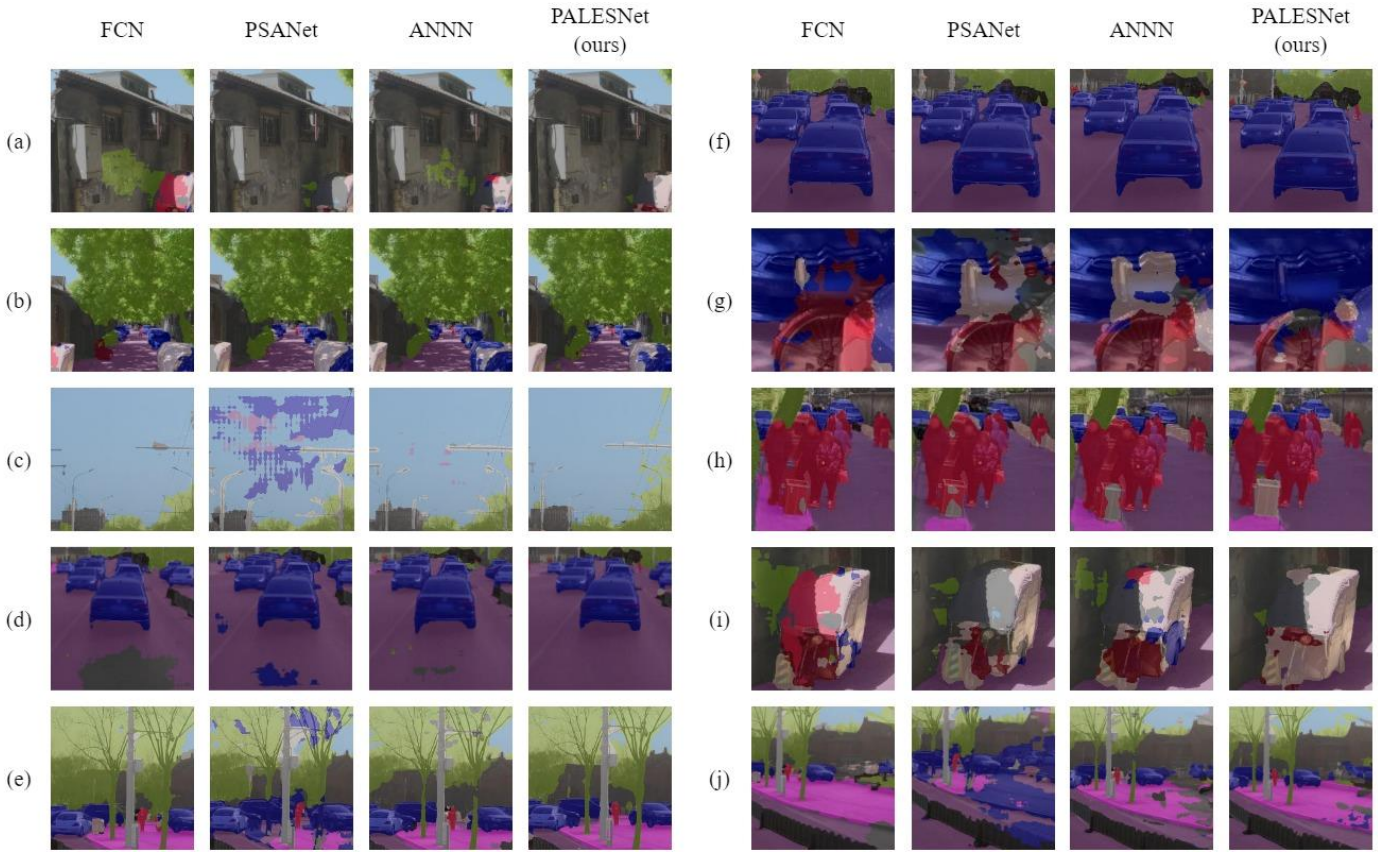

Figure 10. We show the category-by-category segmentation results for (**a**) building, (**b**) vegetation, (**c**) sky, (**d**) road, (**e**) sign, (**f**) motor vehicles, (**g**) non-motor vehicles, (**h**) person, (**i**) clutter, and (**j**) pavements, respectively.

In the modern street scenario, it can be seen that FCN achieved a better performance on the landscape features with a larger area (e.g., road, sky, pedestrian, and vegetation, etc.), but when facing unconventional features (e.g., the incomplete bus on the left), it was prone to failure. The results of PSANet and ANNN were sparse and discontinuous but could recognize unexpected objects. We believe this performance gap was caused by the implementation of the attention mechanism, which can capture the distinguishing features by calculating the relationship between different positions in the image. However, the previous attention-based model cannot handle the complex relationships in such variable situations. Our method, on the other hand, applies the PSA module to fully utilize the resolution both spatial-wise and channel-wise, to distinguish minor differences between every landscape feature.

Regarding the ancient alleyway scenario, this was the trickiest scenario, due to the urban design, complex lighting conditions, and various types of clutter. Apart from the complex physical environment, a challenge for each model also appeared in the limited sample problem, which contains irreplaceable knowledge, and which places a higher demand on feature learning and representation. From the figure above, we can see that the FCN was influenced by the shadow on the wall, and PSANet tended to assign the wrong label (non-motor vehicle) to the tricycle (recognized as clutter because it is illegal). Besides combining the advantage of the multiscale feature extraction ability of the ASPP block and the representation capacity of the PSA attention module, our method enhanced the feature learning ability by imposing and freezing prior knowledge in the network with the transfer learning module. Thus, the proposed method achieved a more robust result in a complex environment.

Apart from the qualitative analysis, we further investigated the segmentation results for each type of landscape feature. As shown in Table 4, we can observe that distinctive features which have a relatively larger area and unifying characteristics in the images (i.e., sky, road, vegetation, and building) had better segmentation results with all of the tested networks compared with the intricate categories, which are fragmentary and diverse (i.e., clutter, person, and pavements). Furthermore, PSANet and ANNN outperformed FCN on some intricate features such as clutter, which proves the potential of the attention mechanism. However, due to the extra complexity, the existing methods could not learn the feature representation with limited samples and this resulted in unsatisfied segmentation results for multiple categories. In contrast, our method achieved the best accuracy in most categories. For the distinctive features, all the models achieved similar results and our method achieved a slight advantage (outperforming by 2–3%). For intricate features, on the other hand, the proposed network outperformed the other methods by a large margin (around 7–10%). These results demonstrate the effectiveness of the PSA module, especially for the reorganization of intricate features.

Table 4. IoUs for each category among the candidate CNNs. Best results in each class are in **bold**; second best are underlined.

Class	FCN	PSANet	ANNN	Our Method
vegetation	74.88	73.93	71.22	**76.83**
sky	**88.12**	70.12	79.91	86.54
road	78.95	76.93	83.81	**86.69**
person	58.07	59.38	55.9	**63.88**
sign	**48.88**	44.02	45.91	46.59
clutter	23.64	27.51	34.15	**35.6**
non-motor vehicles	**50.03**	37.12	38.29	48.47
motor vehicles	56.97	46.34	68.22	**70.2**
building	87.32	86.82	86.77	**89.98**
pavements	53.05	48.58	53.05	**61.42**

We also carried out efficiency analysis experiments on the candidate CNNs, which included the millions of parameters (M), the memory usage (MByte, Mb), the giga floating-point operations per second (GFLOPs), and the infer speed (frames per second, FPS). All the experiments were conducted in the same environment. As shown in Table 5, PSANet is the most complex network with the most parameters (50.57 million), which will cause a negative effect on the feature learning ability in small-sample conditions. The more efficient ANNN (37.66 million parameters), however, cannot fully capture the relationship between similar features. In contrast, the parameters of the proposed PALESNet (43.42 million) were only slightly increased when compared with the conventional FCN (40.93 million), which realizes the full potential of the attention mechanism, while avoiding the side-effects of extra complexity or an excessive compromise on efficiency. The reason for this is that the PSA module will "collapse" the contrary dimension when computing the spatial-wise and channel-wise attention, resulting in lower parameter usage with no performance compromise. Although our method costs more regarding calculation (1619.99 GFlops), we maintain reasonable memory usage (9197 Mb) and infer speed (1.2 FPS) by means of the freezing strategy in the transfer learning module and by only computing both attentions once after the feature extractor.

Table 5. Efficiency analysis result for each method.

Model	Params (M)	Memory (Mb)	GFLOPs	Predict Speed (FPS)
FCN	40.93	8764	1583.94	1.1
PSANet	50.57	10,367	1599.34	1.0
ANNN	37.66	8753	1481.28	1.2
PALESNet (ours)	43.42	9179	1619.99	1.2

4.5. Ablation Study of PALESNet

Our method contained transfer learning, a PSA block, and ASPP block for accurate segmentation in the Beijing Core Area with the proposed HCSV dataset. To further verify the validity of each module in the proposed PALESNet, we designed an ablation experiment, as follows: In the ablation experiment, each module implied in our method was analyzed individually, to check its influence on the whole network. It is worth noting that the FCN was utilized as the baseline, which did not contain a transfer learning module, PSA block, or ASPP block.

In this experiment, the proposed PALESNet had only one more ASPP block compared with the baseline; thus, it could be used to evaluate the effect of ASPP. Each model was trained for 40,000 iterations in the HCSV dataset. As seen in Table 6, the transfer learning could significantly improve the performance of all models, which proved the effectiveness of this technology. More specifically, the mIoU was improved by 17.13%, 12.49%, and 17.24% with FCN, PSANet, and ANNN, respectively. For our method, it was increased by 19.18%, which indicates that our method had the potential to learn more information from the transfer learning procedure. For the ASPP block, it can be seen that our method outperformed the baseline by about 2%, regardless of whether transfer learning was involved or not, which proves the effectiveness of the ASPP block.

Table 6. Ablation study of transfer learning and the ASPP block. Best results in each category are in **bold**.

	Models	***mIoU* (%)**
with-out transfer learning	FCN (baseline)	42.19
	PSANet	42.10
	ANNN	41.86
	PALESNet (only with ASPP)	**44.52**
with transfer learning	FCN (baseline)	59.32
	PSANet	54.59
	ANNN	59.10
	PALESNet (only with ASPP)	**61.90**

As shown in Table 7, we evaluated the performance of the model with and without the attention mechanism. The PSA module improved the mIoU, mAcc, and aAcc of the HCSV dataset by 1.71%, 1.24%, and 0.8%, respectively. This indicates that the PSA block can substantially improve the results, by focusing the identical features for each class and transmitting this information to the subsequent segmentation model.

Table 7. Ablation study of the attention mechanism. Best results in each category are in **bold**.

Model	***mIoU* (%)**	***mAcc* (%)**	***aAcc* (%)**
without PSA	61.99	71.53	89.67
with PSA (our method)	**63.70**	**72.77**	**90.47**

In summary, with the transfer learning technology, PSA block, and ASPP block, our method could make maximum use of the information in the pretrained datasets, as well the limited samples with complex environments in the Beijing Core Area. Therefore, our method could effectively generate a more accurate segmentation map.

5. Discussion

In this section, we first analyze the convergence performance for all comparative networks during the training process, and further discuss the necessity of the proposed HCSV dataset. Finally, we briefly analyze why our method achieved the best performance in the experiment and provide some insights into network design for a small dataset.

- What differed for all comparative methods in the training phase?

The loss and metrics curves in the training phase are shown in Figure 11. As we can see, the original FCN, which did not have the attention-architecture, failed to gain much benefit for the mAcc and aAcc metric in the final stage of training (after 280 epoch). On the other hand, the metrics of the attention-equipped methods continually increased during the whole training process. This may indicate that the attention mechanism could exploit more usable features for the learning of the network. However, the introduction of the attention mechanism also made model more difficult to converge, resulting in jitter of the loss and metric curve. To alleviate this problem, we fused the multi-scale information with the ASPP block and froze a few layers in the network (see Section 3.3.2) during the final training, to stabilize and accelerate the convergence process. As the result, the metric curve of our PALESNet was much smoother than PSANet, and the converge speed of our model was significantly faster than ANNN.

- Is it necessary to create a new dataset for the study in the Historic-Core in Beijing?

As we mentioned before, the existing datasets do not cover unique areas that contain numerous historical monuments, such as in the Beijing Core Area. Thus, the proposed HCSV dataset could provide a novel data source to train and evaluate deep learning models for related studies. As far as we know, we are the first to create a dataset specifically for this historic district in this urban area. The current methods cannot generate suitable results without training on the HCSV dataset. In addition, to contribute to scenery research and preservation in Beijing, it may even be used in other scenarios. The historic architecture shares a similar pattern in China, and models pretrained in the HCSV dataset could converge faster on other historical and cultural blocks.

- What is the best semantic segmentation architecture for the BSV data in Historical-Core?

In the comparison, we found that our method achieved the best performance in both overall, and most class-by-class, accuracy. This outstanding performance may be due to its atrous spatial pyramid pooling structure, which could capture the rich context information and help to understand the complex surface features. The reason for the disappointing performance of the more recent networks with the attention mechanism could be the complexity and diversity of the HCSV data, which may have mislead the networks and resulted in incorrect relationships between various locations.

As for the best semantic segmentation methods for the Historical-Core in Beijing, we believed that this depends on the volume of the annotated images for training. When there are limited amounts of labeled data, a low-capacity network would have a better generalized performance. For example, the Fast-SCNN [47] can output promising segmentation results on Cityscapes without requiring a pre-training process, and the limited training samples will cause less impact compared to the high-capacity DCNNs. However, the attention mechanism-equipped methods, especially the newly emerged transformer-based networks [48,49], cannot learn enough knowledge to distinguish complicated surface features and establish appropriate relationships using the limited annotation samples [50]. Another possible reason for the best performance of the proposed method is that it achieves a balance between network capacity and the amount of data.

Figure 11. Loss and metrics curves for (**a**) FCN, (**b**) PSANet, (**c**) ANNN, and (**d**) PALESNet. It can be seen that compared to the original FCN, the attention-based methods could continually learn information from the training and resulted in the upward trend on all evaluation metrics. Moreover, the proposed PALESNet had a smoother convergence curve than PSANet and had a higher convergence speed compared to ANNN.

6. Conclusions and Future Work

The rapidly evolving online map services with street view images provide a novel perspective to observe the urban landscape and environmental situation, especially for the area with diverse landscape features in the Beijing Core Area. To address the question of data shortage, this paper provides a small yet comprehensive History-Core Street View dataset for related research. Furthermore, we proposed a DNN-based method with transform learning technology, a PSA attention block, and a ASPP block to perform an accurate landscape features extraction for Chinese historic districts. To alleviate the negative effects caused by the limited sample problem, the transfer learning module can transfer the knowledge from existing datasets to the proposed network, to assist in discriminating identical features in the HCSV dataset. Moreover, the PSA attention block can distinguish intricate features, whereas the ASPP block can extract multi-scale features and, thus, can help the model extract landscape features more accurately, especially in a complex environment. Compared to other state-of-the-art methods, i.e., ANNN and PSANet, our network achieved the highest accuracy, with an mIoU of 63.7% on the HCSV dataset.

In the future, we will further explore recent weakly-supervised and transformer technology and develop effective landscape feature extraction methods that can distinguish more types of features with a higher accuracy under a complex environment and small sample situation. In addition, the proposed method could also be promoted to other historic districts in China such as the Ancient Town of Fenghuang and Lilong in Shanghai, to support the protection of traditional landscapes by providing land feature investigation data to the relevant departments.

Author Contributions: Conceptualization, Siming Yin and Xian Guo; methodology, Siming Yin; validation, Siming Yin; formal analysis, Siming Yin; resources, Jie Jiang; data curation, Siming Yin; writing—original draft preparation, Siming Yin; writing—review and editing, Xian Guo and Jie Jiang; supervision, Xian Guo; project administration, Jie Jiang; funding acquisition, Jie Jiang. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by The National Key R&D Program of China (2021YFE0117500), The Pyramid Talent Training Project of Beijing University of Civil Engineering and Architecture (JDYC20200322), The Fundamental Research Funds for Beijing University of Civil Engineering and Architecture (X20044).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Shan, J.X. *Conservation of Historic and Cultural Districts*; Tianjin University Press: Tianjin, China, 2015. (In Chinese)
2. Cai, X.F. Analysis and Regulation on City Style and Feature. Ph.D. Thesis, Tongji University, Shanghai, China, 2006; ISBN 978-7-5618-5303-0. (In Chinese)
3. Mangi, M.Y.; Yue, Z.; Kalwar, S.; Ali Lashari, Z. Comparative analysis of urban development trends of Beijing and Karachi metropolitan areas. *Sustainability* **2020**, *12*, 451. [CrossRef]
4. Wherrett, J.R. Creating landscape preference models using internet survey techniques. *Landsc. Res.* **2000**, *25*, 79–96. [CrossRef]
5. Vermeulen, F.; Keay, S.J.; Burgers, G.-J.; Corsi, C. *Urban Landscape Survey in Italy and the Mediterranean*; Oxbow Books: Oxford, UK, 2012; ISBN 9781842174869.
6. Ahern, J. Urban landscape sustainability and resilience: The promise and challenges of integrating ecology with urban planning and design. *Landsc. Ecol.* **2013**, *28*, 1203–1212. [CrossRef]
7. Li, Z.; Han, X.; Lin, X.; Lu, X. Quantitative analysis of landscape efficacy based on structural equation modelling: Empirical evidence from new Chinese style commercial streets. *Alex. Eng. J.* **2021**, *60*, 261–271. [CrossRef]
8. Liu, Y.; Wang, R.; Lu, Y.; Li, Z.; Chen, H.; Cao, M.; Zhang, Y.; Song, Y. Natural outdoor environment, neighbourhood social cohesion and mental health: Using multilevel structural equation modelling, streetscape and remote-sensing metrics. *Urban For. Urban Green.* **2020**, *48*, 126576. [CrossRef]

9. Zhang, X. Practice teaching of landscape survey course based on ecognition remote sensing image interpretation* technology. *Educ. Sci. Theory Pract.* **2018**, *18*, 1411–1423. [CrossRef]
10. Tang, J.; Long, Y. Measuring visual quality of street space and its temporal variation: Methodology and its application in the Hutong area in Beijing. *Landsc. Urban Plan* **2019**, *191*, 103436. [CrossRef]
11. Xu, Z.; Wu, Y.; Lu, X.Z.; Jin, X.L. Photo-realistic visualization of seismic dynamic responses of urban building clusters based on oblique aerial photography. *Adv. Eng. Inform.* **2020**, *43*, 17. [CrossRef]
12. Ravindran, R.; Santora, M.J.; Jamali, M.M. Multi-Object Detection and Tracking, Based on DNN, for Autonomous Vehicles: A Review. *IEEE Sens. J.* **2021**, *21*, 5668–5677. [CrossRef]
13. Gong, F.Y.; Zeng, Z.C.; Zhang, F.; Li, X.J.; Ng, E.; Norford, L.K. Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. *Build. Environ.* **2018**, *134*, 155–167. [CrossRef]
14. Liang, J.; Gong, J.; Sun, J.; Zhou, J.; Li, W.; Li, Y.; Liu, J.; Shen, S. Automatic sky view factor estimation from street view photographs—A big data approach. *Remote Sens.* **2017**, *9*, 411. [CrossRef]
15. Cheng, L.; Chu, S.S.; Zong, W.W.; Li, S.Y.; Wu, J.; Li, M.C. Use of Tencent Street View Imagery for Visual Perception of Streets. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 265. [CrossRef]
16. Rundle, A.G.; Bader, M.D.; Richards, C.A.; Neckerman, K.M.; Teitler, J.O. Using Google Street View to audit neighborhood environments. *Am. J. Prev. Med.* **2011**, *40*, 94–100. [CrossRef] [PubMed]
17. Li, X.J.; Ratti, C.; Seiferling, I. Quantifying the shade provision of street trees in urban landscape: A case study in Boston, USA, using Google Street View. *Landsc. Urban Plan.* **2018**, *169*, 81–91. [CrossRef]
18. Li, X.J.; Zhang, C.R.; Li, W.D. Building block level urban land-use information retrieval based on Google Street View images. *GIScience Remote Sens.* **2017**, *54*, 819–835. [CrossRef]
19. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
21. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
22. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.
23. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
24. Huo, X.; Xie, L.; He, J.; Yang, Z.; Zhou, W.; Li, H.; Tian, Q. ATSO: Asynchronous teacher-student optimization for semi-supervised image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 1235–1244.
25. Wang, H.; Zhu, Y.; Adam, H.; Yuille, A.; Chen, L.-C. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 5463–5474.
26. Yuan, X.H.; Shi, J.F.; Gu, L.C. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2021**, *169*, 14. [CrossRef]
27. Yan, Y.L.; Ryu, Y. Exploring Google Street View with deep learning for crop type mapping. *Isprs J. Photogramm. Remote Sens.* **2021**, *171*, 278–296. [CrossRef]
28. Zhang, F.; Wu, L.; Zhu, D.; Liu, Y. Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. *ISPRS J. Photogramm. Remote Sens.* **2019**, *153*, 48–58. [CrossRef]
29. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
30. Middel, A.; Lukasczyk, J.; Zakrzewski, S.; Arnold, M.; Maciejewski, R. Urban form and composition of street canyons: A human-centric big data and deep learning approach. *Landsc. Urban Plan* **2019**, *183*, 122–132. [CrossRef]
31. Ye, Y.; Zeng, W.; Shen, Q.M.; Zhang, X.H.; Lu, Y. The visual quality of streets: A human-centred continuous measurement based on machine learning algorithms and street view images. *Environ. Plan. B Urban Anal. City Sci.* **2019**, *46*, 1439–1457. [CrossRef]
32. Suel, E.; Bhatt, S.; Brauer, M.; Flaxman, S.; Ezzati, M. Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. *Remote Sens. Environ.* **2021**, *257*, 11. [CrossRef] [PubMed]
33. Zhang, L.Y.; Pei, T.; Wang, X.; Wu, M.B.; Song, C.; Guo, S.H.; Chen, Y.J. Quantifying the Urban Visual Perception of Chinese Traditional-Style Building with Street View Images. *Appl. Sci.* **2020**, *10*, 5963. [CrossRef]
34. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 June 2022.
35. Zhu, Z.; Xu, M.; Bai, S.; Huang, T.; Bai, X. Asymmetric non-local neural networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 593–602.

36. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, Florida, USA, 20–25 July 2009; pp. 248–255.
37. Brostow, G.J.; Fauqueur, J.; Cipolla, R. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognit. Lett.* **2009**, *30*, 88–97. [CrossRef]
38. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3213–3223.
39. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
40. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2636–2645.
41. Huang, X.; Cheng, X.; Geng, Q.; Cao, B.; Zhou, D.; Wang, P.; Lin, Y.; Yang, R. The apolloscape dataset for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 954–960.
42. Semantic Segmentation Editor Contributors. Semantic Segmentation Editor. Available online: https://github.com/Hitachi-Automotive-And-Industry-Lab/semantic-segmentation-editor (accessed on 18 May 2022).
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, USA, 3–6 December 2012; pp. 1097–1105.
45. Veit, A.; Wilber, M.J.; Belongie, S. Residual networks behave like ensembles of relatively shallow networks. In Proceedings of the 30th Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 550–558.
46. MMSegmentation Contributors. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 18 May 2022).
47. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-scnn: Fast semantic segmentation network. *arXiv* **2019**, arXiv:1902.04502.
48. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 3–7 May 2021.
49. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021.
50. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 January 2022.