remote sensing

MDPI

Article

The Cost of Urban Renewal: Annual Construction Waste Estimation via Multi-Scale Target Information Extraction and Attention-Enhanced Networks in Changping District, Beijing

Lei Huang [1], Shaofu Lin [1], Xiliang Liu [1,*], Shaohua Wang [2], Guihong Chen [3], Qiang Mei [4] and Zhe Fu [5]

[1] Faculty of Information Technology, Beijing University of Technology, Chaoyang District, Beijing 100124, China; huang-lei@emails.bjut.edu.cn (L.H.); linshaofu@bjut.edu.cn (S.L.)
[2] State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; wangshaohua@aircas.ac.cn
[3] Beijing Big Data Centre, Chaoyang District, Beijing 100101, China; chengh@jxj.beijing.gov.cn
[4] Navigation College, Jimei University, Xiamen 361021, China; meiqiang@jmu.edu.cn
[5] Administrative Examination and Approval Bureau of the Beijing Economic-Technological Development Area, Beijing 100176, China; fuzhe@bda.gov.cn
* Correspondence: liuxl@bjut.edu.cn

Abstract: Construction waste is an inevitable byproduct of urban renewal, causing severe pressure on the environment, health, and ecology. Accurately estimating the production of construction waste is crucial for assessing the consumption of urban renewal. However, traditional manual estimation methods rely heavily on statistical data and historical experience, which lack flexibility in practical applications and are time-consuming and labor-intensive. In addition, their accuracy and timeliness need to be improved urgently. Fortunately, with the advantages of high-resolution remote sensing images (HRSIs) such as strong timeliness, large amounts of information, and macroscopic observations, they are suitable for the large-scale dynamic change detection of construction waste. However, the existing deep learning models have a relatively poor ability to extract and fuse features for small and multi-scale targets, and it is difficult to deal with irregularly shaped and fragmented detection areas. Therefore, this study proposes a Multi-scale Target Attention-Enhanced Network (MT-AENet), which is used to dynamically track and detect changes in buildings and construction waste disposal sites through HRSIs and accurately estimate the annual production of urban construction waste. The MT-AENet introduces a novel encoder–decoder architecture. In the encoder, ResNet-101 is utilized to extract high-level semantic features. A depthwise separable-atrous spatial pyramid pooling (DS-ASPP) module with different dilation rates is constructed to address insufficient receptive fields, resolving the issue of discontinuous holes when extracting large targets. A dual-attention mechanism module (DAMM) is employed to better preserve positional and channel details. In the decoder, multi-scale feature fusion (MS-FF) is utilized to capture contextual information, integrating shallow and intermediate features of the backbone network, thereby enhancing extraction capabilities in complex scenes. The MT-AENet is used to extract buildings and construction waste at different periods in the study area, and the actual production and landfill volume of construction waste are calculated based on area changes, indirectly measuring the rate of urban construction waste resource conversion. The experimental results in Changping District, Beijing demonstrate that the MT-AENet outperforms existing baseline networks in extracting buildings and construction waste. The results of this study are validated according to government statistical standards, providing a promising direction for efficiently analyzing the consumption of urban renewal.

Keywords: high-resolution remote sensing image segmentation; attention-enhanced network; building extraction; construction waste extraction; construction waste disposal; urban renewal

Citation: Huang, L.; Lin, S.; Liu, X.; Wang, S.; Chen, G.; Mei, Q.; Fu, Z. The Cost of Urban Renewal: Annual Construction Waste Estimation via Multi-Scale Target Information Extraction and Attention-Enhanced Networks in Changping District, Beijing. *Remote Sens.* **2024**, *16*, 1889. https://doi.org/10.3390/rs16111889

Academic Editors: Wen Yang, Pengyuan Lv, Chen Wu and Naoto Yokoya

Received: 10 April 2024
Revised: 21 May 2024
Accepted: 23 May 2024
Published: 24 May 2024

1. Introduction

Urban renewal [1] typically involves the demolition of old buildings, construction of new structures, and land re-planning. This comprehensive approach effectively addresses land shortages and optimizes urban spatial layouts and functional designations. Influenced by factors such as urban pollution [2], land resources [3], and regional policies, the demolition and transformation of old buildings have become crucial methods to meet the requirements of urban renewal. In particular, in megacities, there is an urgent need for urban renewal, which is a key development goal for urban planning, the establishment of smart cities, and sustainable growth [4].

Construction waste is an inevitable byproduct of the urban renewal process. The extensive and large-scale urban construction has significantly increased construction waste [5], causing severe environmental pollution, human health risks, and ecological pressure [6]. The European Union (EU) construction industry generates over 500 million tons of construction waste annually, accounting for 50% of all waste produced in the EU [7]. In the United States, the construction industry produces 700 million tons of construction waste each year [8]. Similarly, China faces formidable challenges, with its urbanization rate escalating at an unprecedented pace, surging from 51.83% in 2011 to 65.22% in 2022 [9]. Construction waste constitutes over 40% of urban solid waste [10]. With the rapid development of urbanization in China over the past decade, from 2006 to 2020, construction waste production has surged from 470 million tons to a staggering 3037 million tons. It reached 3209 million tons in 2021 and is projected to surpass 4000 million tons by 2026 [11]. Therefore, there is an urgent need to accurately estimate the annual production of construction waste and implement appropriate waste management to improve urban quality, addressing the consumption issues in the urban renewal process [12].

Numerous methods currently exist for estimating the annual production of construction waste. Wu et al. [13] classify the methods for estimating construction waste into six types, including site visit (SV), generation rate calculation (GRC), lifetime analysis (LA), classification system accumulation (CSA), variable modeling (VM), and other methods. (1) The SV method requires on-site investigations, including direct measurements through weight and volume [14] or indirect measurements using other easily accessible indicators (such as waste transport tickets) [15]. Direct measurements can provide the most accurate waste yield but consume a significant amount of time, money, and labor. Indirect measurements can only roughly reflect the waste generation. (2) The GRC method calculates the total waste volume by multiplying the quantity of specific units by the corresponding generation rate. The per capita multiplier [16] is the earliest quantification method for construction and demolition waste in GRC, offering a simple method for quantifying waste from construction and demolition activities in an area. However, economic conditions can lead to significant fluctuations in construction and demolition activities while the population remains almost constant. (3) The LA method assumes that all buildings must be demolished after a certain lifespan and infers the construction waste production by calculating the sum of weights to be demolished at expiration [17]. However, this method requires appropriate assumptions about the lifespan of buildings and cannot provide detailed volume estimates at the material level. (4) The CSA method combines the GRC method with waste classification, offering a more detailed waste estimate by quantifying each specific material [18]. However, this method requires region-specific data, which may not be suitable for industries with different construction technologies. (5) The VM method designs predictive models such as multiple linear regression and gray prediction models based on accessible variables [19]. However, such methods are generally only suitable for short-term forecasting and cannot accurately estimate the annual production of construction waste. (6) Other methods mainly include the popular shallow machine learning (SML) methods for prediction in recent years [20,21]. However, the traditional methods consume considerable manpower, resources, and time and lack high accuracy and efficiency.

In recent years, with the continuous advancement of remote sensing technology and the widespread application of deep learning in target extraction, estimating annual construction waste production at a macroscopic level using HRSIs has become a crucial focus of current research. HRSIs possess advantages such as high spatial resolution, vital timeliness, and abundant information [22], making them suitable for large-scale macroscopic observations of changes in waste piles. However, construction waste consists of debris from the construction process and demolition waste from dismantling activities [23]. Except for the recycled waste, the remaining debris is sent to construction waste disposal sites for treatment. Hence, accurately estimating the generation of urban construction waste solely based on changes in waste piles at construction waste disposal sites is challenging. It is essential to comprehensively consider variations in both building areas and construction waste areas depicted in the images. Typically, the methods involve calculating construction waste production based on area multiplied by generation rate [24,25], subcategorizing engineering and demolition waste, and tracking changes in construction waste disposal sites. These approaches help eliminate errors caused by intermediate variables, enabling the analysis of urban construction waste landfill volume and resource conversion. This, in turn, enhances the precision of estimating both the production of urban construction waste and the capacity for its disposal.

The change in building area calculation requires building identification as the initial step. With the introduction of convolutional neural networks (CNNs) [26], numerous works have emerged [27]. For instance, Kang et al. [28] proposed an efficient end-to-end fully convolutional model, EU-Net, designed for extracting buildings from optical remote sensing images. Shao et al. [29] introduced a building residual refine network (BRRNet) based on an encoder–decoder structure, enabling the accurate and comprehensive extraction of complex-shaped buildings. Similarly, He et al. [30] presented an automated building extraction method utilizing fully convolutional networks (FCNs) and conditional random fields (CRFs), which are effectively applicable for building target extraction in remote sensing images. Chen et al. [31] employed the encoder–decoder backbone of DeepLabV3+. They proposed a dense residual neural network (DR-Net) by combining densely connected convolutional neural network (DCNN) and residual network (ResNet) structures, demonstrating efficient building extraction. Wang et al. [32] combined a UNet, residual learning, atrous spatial pyramid pooling, and focal loss to propose an effective model, the residual UNet (RU-Net), for building extraction, achieving three to four times higher efficiency compared to FastFCNs and DeepLabV3+. Inspired by graph convolutional networks (GCNs), which can naturally model long-range spatial relations in HRSIs [33], it has been observed that multi-scale feature fusion enhances the accuracy of building identification. For instance, Wang et al. [34] proposed an end-to-end multi-scale boundary detection network (MBDNet) that combines a multi-level neural network structure with a boundary detector, improving building extraction through boundary perception. Zhang et al. [35] proposed the dual spatial attention transformer net (DSAT-Net), a high-precision building extraction model, and designed an efficient dual spatial attention transformer (DSAFormer) to address the shortcomings of the standard vision transformer. Zhang et al. [36] proposed the shunted dual skip connection UNet (SDSC-UNet), which introduces a novel shunted transformer to enable the model to establish global dependencies while capturing multi-scale information internally and is designed for high-precision building extraction.

Tracking and identifying construction waste disposal sites through high-resolution images has undergone significant research development. For example, Chen et al. [37] proposed an optimal method that combines morphological indices and hierarchical segmentation. This approach enhanced the separability of construction waste from surrounding ground objects by comparing differences in spectral, geometric shape, and texture aspects, effectively addressing the spectral confusion between construction waste and the surrounding ground. Similarly, Davis et al. [38] designed a deep convolutional neural network to simulate a real construction site scenario, which is challenging to classify on site. They employed digital images of waste deposited in construction site bins (artificial artifacts) to

identify seven types of construction waste. Sun et al. [39] addressed the issue of insufficient data for solid waste detection by introducing a data augmentation strategy. They proposed an improved pix2pix model to generate sufficient high-quality synthetic images for solid waste detection, thereby establishing a landfill dataset. Li et al. [40] constructed a new solid waste detection dataset and proposed a location-guided key point network with multiple enhancements (LKN-ME) for urban solid waste detection. Xiaoyu et al. [41] used the DeeplabV3+ network model and encoder to locate shallow and high-level semantic features. This enabled them to identify the location, type, area, and volume of illegally accumulated construction waste in remote sensing images. Zhang et al. [42] proposed the ConvLSR-Net, where they appended a novel efficient vision transformer module called long-short-range transformer (LSRFormer) to learn both local and global features at each stage, making it applicable for the semantic segmentation extraction of various types of aerial images and construction waste in complex scenes.

The studies above indicate notable advancements in applying deep learning to identify buildings and construction waste in HRSIs. However, the following issues still exist in the current research:

1. Challenges in building recognition within complex scenes often arise from small targets, diverse sizes, varied shapes, and different types of buildings. These factors contribute to a relatively lower accuracy in building recognition and suboptimal image segmentation [43]. The inadequate fusion of multi-scale features may lead to misclassifications [44]. Traditional convolutions frequently struggle to retain spatial details effectively, leading to blurred boundaries and overlooking small buildings. Fixed receptive fields consistently lead to discontinuous gaps when extracting information from large buildings [45].
2. Complex scenes present multiple challenges in identifying construction waste. These challenges include the following: (1) Construction waste is often situated in environments with similar feature information. In such scenarios, the network encounters difficulty in emphasizing less prominent characteristics. (2) Construction waste typically exhibits irregular shapes, fragmentation, and dispersion, posing challenges to the network's capacity to capture spatial information and impeding accurate detection. (3) The unique attributes of construction waste, including color, shape, and texture, combined with substantial distinctions from other ground objects in satellite image backgrounds, make simple CNN structures insufficient for identifying construction waste in complex environments. A deeper and more specific network architecture is essential to accurately delineate construction waste areas in remote sensing images with complex backgrounds [46]. To mitigate labor and time costs and enhance the efficiency of estimating construction waste production, there is an urgent need to develop a flexible multi-scale target identification model addressing construction waste subdivision identification and tracking changes in construction waste disposal sites.
3. The lack of a dataset for construction waste disposal sites constitutes a significant issue. There is a severe shortage of publicly available datasets designed explicitly for construction waste identification, and the existing datasets adhere to different standards. The commonly used datasets for construction waste extraction include the AerialWaste dataset [47] and the SWAD dataset [48]. Still, they exhibit several shortcomings in practical applications: (1) The AerialWaste dataset lacks the annotated information required for semantic segmentation, as its classification is based on the presence of solid waste rather than on each pixel in the image. This limitation impedes the quantitative analysis of waste production. (2) The SWAD dataset employs satellite spatial resolutions of 180 cm of GSD (Ground Sampling Distance), which is not a sub-meter satellite image. Consequently, the images fail to offer clear views of construction waste disposal sites, complicating the discernment of typical features. (3) Owing to indistinct details, the color of construction waste may slightly differ from its surroundings. However, these differences remain imperceptible due to the lower

resolution, resulting in unclear heights and shapes of waste piles and an inability to accurately estimate activity at waste disposal sites.

In response to the abovementioned challenges, a multi-scale target attention-enhanced network (MT-AENet) is introduced to extract buildings and construction waste from complex backgrounds through semantic segmentation in HRSIs. The main contributions of this study can be summarized as follows:

(1) A novel model, the MT-AENet, based on an encoder–decoder structure, is designed explicitly for feature extraction in HRSIs. The encoder utilizes ResNet101 as the backbone network to extract high-dimensional features. The DS-ASPP and multi-scale feature fusion modules are integrated into the MT-AENet to extract features from both local and global image levels, which helps to fuse contextual information better. The dual-attention mechanism module further improves the accuracy and efficiency of detecting buildings and construction waste in HRSIs with complex backgrounds.

(2) A method for calculating the annual production of construction waste based on analyzing building changes using remote sensing images is proposed. By leveraging deep learning algorithms to extract and analyze the distribution of buildings in the same area at different times, the area changes of newly added and demolished buildings can be quickly and accurately obtained, and construction waste production can be accurately estimated. Additionally, by analyzing the changes in the landfill volume of construction waste disposal sites and assessing the landfill volume and resource conversion capacity, the annual disposal capacity of urban construction waste can be obtained.

(3) A comprehensive dataset, namely, the Construction Waste Disposal Sites Dataset (CWDSD), has been established for the internal area identification of construction waste disposal sites. Taking Changping and Daxing District, Beijing, China as an example, this dataset of construction waste disposal sites is curated from the sub-meter level satellite GF-2 and Google Earth. Detailed labeled images are provided, annotating various areas within the disposal site, including vacant landfills, engineering facilities, and waste storage areas.

The remainder of this study is organized as follows. Section 2 provides the details of the proposed method. Section 3 describes the dataset and experimental setup. Section 4 introduces the experimental results, comparing the MT-AENet to traditional and current state-of-the-art methods. A detailed calculation of the construction waste yield is presented. Section 5 discusses the differences between the existing work and our approach. Section 6 summarizes the conclusions and future research.

2. Methodology

This section provides a detailed description of the method for accurately estimating the annual production of urban construction waste using deep learning algorithms based on remote sensing images. The overall system architecture is illustrated in Figure 1. Initially, the raw GF-2 satellite images undergo preprocessing, including radiometric calibration, atmospheric correction, and orthorectification. A sample library containing labeled images of buildings and construction waste disposal sites is established through manual annotation. An MT-AENet model based on the encoder–decoder structure is constructed, and features are extracted separately for both the buildings and internal areas of construction waste disposal sites. By integrating two years of image data, each region's building area changes are determined, and combined with the relevant generation rates, the annual production of construction waste (including engineering and demolition waste) is estimated. Considering that apart from recycled construction waste, the remaining waste will be landfilled in construction waste disposal sites, comparing the actual construction waste and the additional waste in the landfill can yield the urban construction waste resource conversion rate.

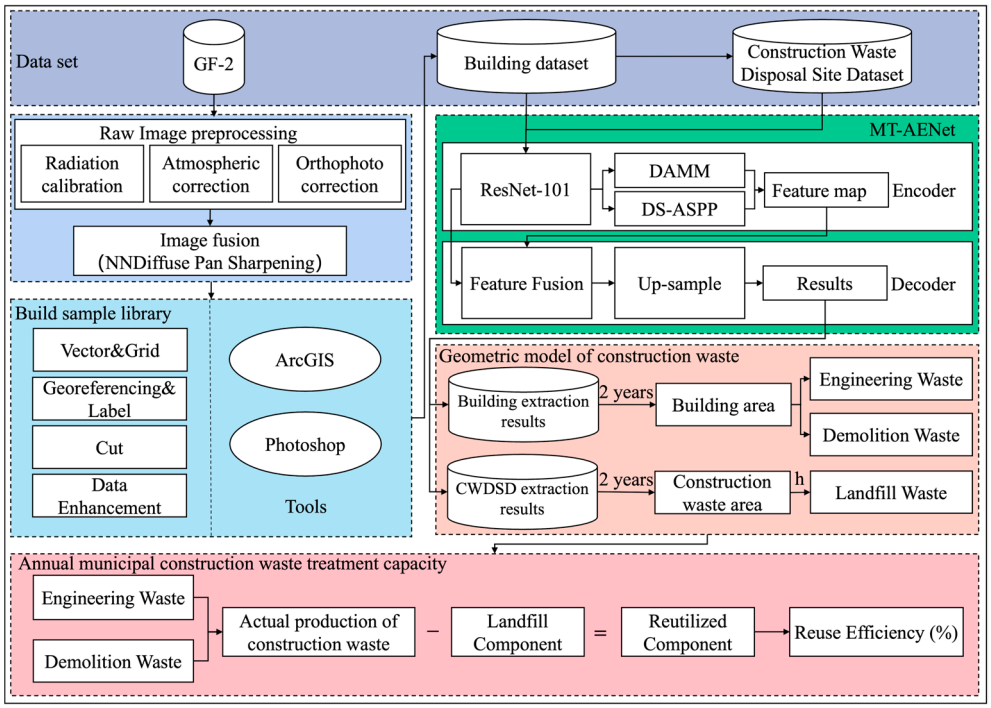


Figure 1. The overall system architecture diagram. It depicts the utilization of GF-2 remote sensing images in conducting experiments related to the extraction of buildings and construction waste disposal sites, as well as precise estimation of the annual production of construction waste.

2.1. MT-AENet Structure

In this section, the overall framework of the MT-AENet is introduced, including a backbone network composed of RestNet-101, an encoder incorporating the dual-attention mechanism module (DAMM) and the depthwise separable-atrous spatial pyramid pooling (DS-ASPP) module, and a decoder integrating a feature fusion module for multi-scale feature fusion. The network structure of the MT-AENet is shown in Figure 2.

2.1.1. Encoder

Utilizing ResNet-101 as the backbone network for the encoder (see the encoder section in Figure 2), deep convolution is performed through convolutional and pooling layers. Adjusting the stride of the convolutional layers in Layer 4 to 1 prevents the downsizing of the feature map, thus enhancing the resolution of the output feature map. This aids in preserving more detailed information and extracting denser features. The final feature map obtained is 1/16 the size of the original image, and experimental results demonstrate that the optimal trade-off between speed and accuracy is achieved when the encoder's output stride is 16. The DS-ASPP module comprises a 1×1 convolution, three 3×3 depthwise separable dilated convolutions with 6, 12, and 18 dilation rates, and a global pooling operation. A dual-attention mechanism composed of positional and channel attention is introduced, connected in parallel with the DS-ASPP module, to extract multi-scale contextual information and enhance the robustness of the whole network. In the encoder structure, the input image passes through the ResNet-101 backbone network, obtaining low-level and high-level feature maps. The low-level feature map enters the decoder layer, while the high-level feature maps go through the DAMM and the DS-ASPP module. The results are summed element-wise to obtain the output of the encoder layer, which then serves as the input for the decoder layer.

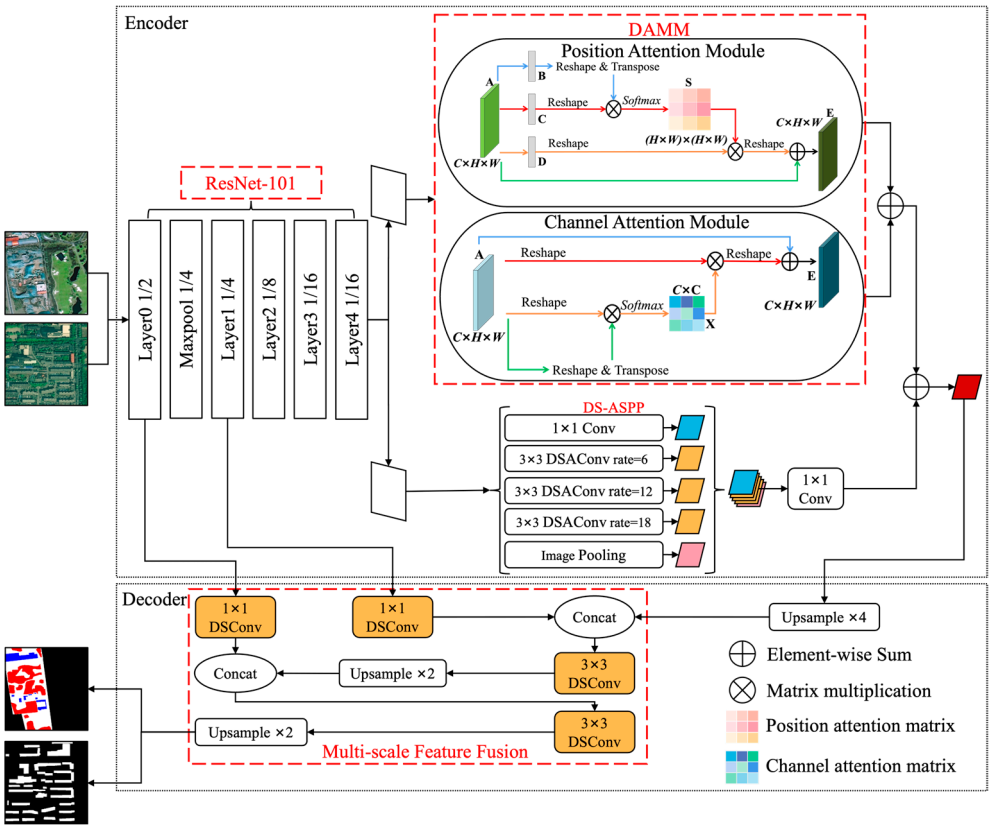


Figure 2. The MT-AENet network structure with the DS-ASPP module and DAMM. In the encoding stage, the ResNet-101 is the backbone network connected to the DS-ASPP module and DAMM. In the decoding stage, the image's spatial resolution and positional information are recovered by up-sampling and feature fusion.

2.1.2. Decoder

The decoder (see the decoder section in Figure 2) restores the image's spatial resolution and positional information. In the decoding stage, the feature map obtained in the encoding stage undergoes four-fold bilinear interpolation up-sampling. It is fused with the intermediate feature map generated by Layer 1 of the encoder's backbone network. A feature fusion module is added to combine low-level and intermediate features of the backbone network. The result undergoes 3 × 3 convolution, followed by two-fold bilinear interpolation up-sampling, merging with the feature map generated by Layer 0 of the backbone network after a 1 × 1 convolution, enhancing the network's ability to acquire multi-scale contextual features. A 3 × 3 convolution is applied to the feature map, followed by two-fold bilinear interpolation up-sampling, yielding an ultimately finer segmentation result.

2.2. Detailed Optimization of MT-AENet

This section outlines specific enhancements to the MT-AENet to improve its capacity for extracting buildings and construction waste in complex scenes. These modifications address issues such as insufficient multi-scale feature fusion, blurred boundaries, and high computational complexity in image recognition.

2.2.1. Feature Extraction Backbone Network ResNet-101

In multi-scale feature recognition, shallow networks can only extract basic features like image edges, colors, and textures. To enhance classification accuracy by extracting more features, deepening the network to capture abstract features is necessary. However, blindly increasing the depth of convolutional neural networks can lead to problems such as gradient explosion or vanishing, making the model challenging to train and causing a decrease

in accuracy. In 2016, He et al. [49] introduced the ResNet (Residual Neural Network), addressing this issue by introducing residual connections. The ResNet introduces cross-layer jump connections that add the previous layer's output directly to the subsequent layer's output, avoiding the problem of vanishing gradients.

The ResNet-101, part of the ResNet series, is a deep convolutional neural network. Compared to shallower networks like the ResNet-34 and the ResNet-50, the ResNet-101 has a deeper structure and convolutional kernels, enabling better capture of deep image features, enhancing model performance, and exhibiting superior generalization capabilities. Therefore, in this study, the ResNet-101 was selected as the backbone network for feature extraction in the MT-AENet encoder (see the image input encoder section in Figure 2).

2.2.2. Position Attention Module

The purpose of the positional attention module is to enhance the association of any two points in the image, which can simulate the rich contextual information between the global features so that the same features at different locations can enhance each other and improve the semantic segmentation ability (see the upper branch section of the DAMM in Figure 2). The detailed structure of the positional attention module is shown in Figure 3.

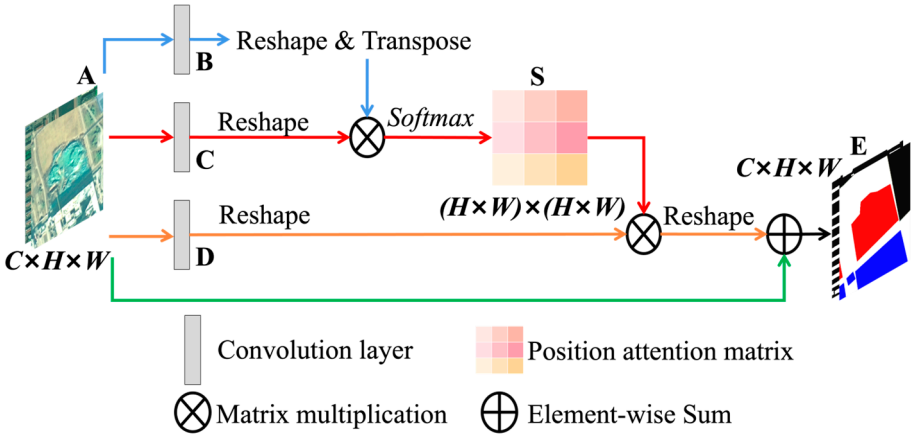


Figure 3. The detailed structure of positional attention.

A local feature $\boldsymbol{A} \in \mathbb{R}^{C \times H \times W}$ matrix can be obtained from the backbone network, and the feature $\boldsymbol{A}$ is convolved by a 1 × 1 convolution operation to obtain the $\boldsymbol{B}$ vector with channel C/8, the $\boldsymbol{C}$ vector with channel C/8, and the $\boldsymbol{D}$ vector with channel C. Next, the matrix $\boldsymbol{B}$ dimension is converted to $\boldsymbol{B'}$, where $\boldsymbol{B'} \in \mathbb{R}^{C \times N}$, N = H × W, and the matrix $\boldsymbol{B'}$ is transposed to $\boldsymbol{B^T}$. The $\boldsymbol{C}$ dimension is converted to $\boldsymbol{C'}$. Matrix multiplication of $\boldsymbol{C'}$ and $\boldsymbol{B^T}$ is performed to obtain the matrix (H × W) × (H × W). Finally, the positional attention matrix $\boldsymbol{S} \in \mathbb{R}^{N \times N}$ is computed using SoftMax. The obtained matrix $\boldsymbol{S}$ can be regarded as the weights obtained through the positional attention mechanism. For the input local feature information, $\boldsymbol{A}$, the positional attention $\boldsymbol{S}$ is obtained through two-dimensional transformations and the SoftMax operation, which is calculated as follows:

$$S_{ji} = \frac{exp(B_i \cdot C_j)}{\sum_{i=1}^{N} exp(B_i \cdot C_j)}. \quad (1)$$

where B_i represents the i-th element in matrix $\boldsymbol{B} \in \mathbb{R}^{N \times C}$, C_j represents the j-th element in matrix $\boldsymbol{C} \in \mathbb{R}^{C \times N}$, N represents the number of elements in the current channel, and S_{ji} denotes the influence factor of the i-th position on the j-th position.

The final result, denoted as $\boldsymbol{E}$, is obtained by multiplying the matrix $\boldsymbol{D}$ with the positional attention matrix through dimensional transformation and then summing the

elements with the original feature map $\boldsymbol{A}$, where $\boldsymbol{E} \in \mathbb{R}^{C \times H \times W}$. The computation formula is as follows:

$$E_j = \alpha \sum_{i=1}^{N} \left(S_{ji} \cdot D_i \right) + A_j. \tag{2}$$

where $\boldsymbol{D_i}$ denotes the i-th element in the matrix $\boldsymbol{D} \in \mathbb{R}^{N \times C}$; S_{ji} is the i-th positional element of the matrix $\boldsymbol{S}$; and α is a learnable parameter, which is initially zero.

From formula (2), the final feature $\boldsymbol{E}$ of each location is the weighted sum of all the location features and the original feature; so, the positional attention mechanism has a global context view and tries to selectively aggregate contexts based on the positional attention so that similar semantic features promote each other and maintain semantic consistency.

2.2.3. Channel Attention Module

The channel attention mechanism is widely used in deep learning. As there are different connections between different channel feature maps, the channel attention mechanism can highlight the feature maps connected by extracting the semantic information between different channels with various weights (see the lower branch section of the DAMM in Figure 2). Unlike the position attention module, the channel attention module computes the channel attention matrix $\boldsymbol{X} \in \mathbb{R}^{C \times C}$ directly from $\boldsymbol{A} \in \mathbb{R}^{C \times H \times W}$. The feature matrix $\boldsymbol{A}$ dimensions are converted into $\boldsymbol{A'} \in \mathbb{R}^{C \times N}$, where N = H × W. The matrix multiplication of $\boldsymbol{A'}$ with the transposed matrix of $\boldsymbol{A'}$ and passing it through the SoftMax layer yields $\boldsymbol{X}$, where the maximum value is subtracted from each row in $\boldsymbol{X}$ to increase the focus on other similarities. The obtained attention weights are matrix-multiplied with $\boldsymbol{A'}$ and multiplied by a learnable coefficient β. The final result, $\boldsymbol{E}$, is obtained by element-wise summation with the original feature map $\boldsymbol{A}$, where $\boldsymbol{E} \in \mathbb{R}^{C \times H \times W}$. The structure of the channel attention module is shown in Figure 4.

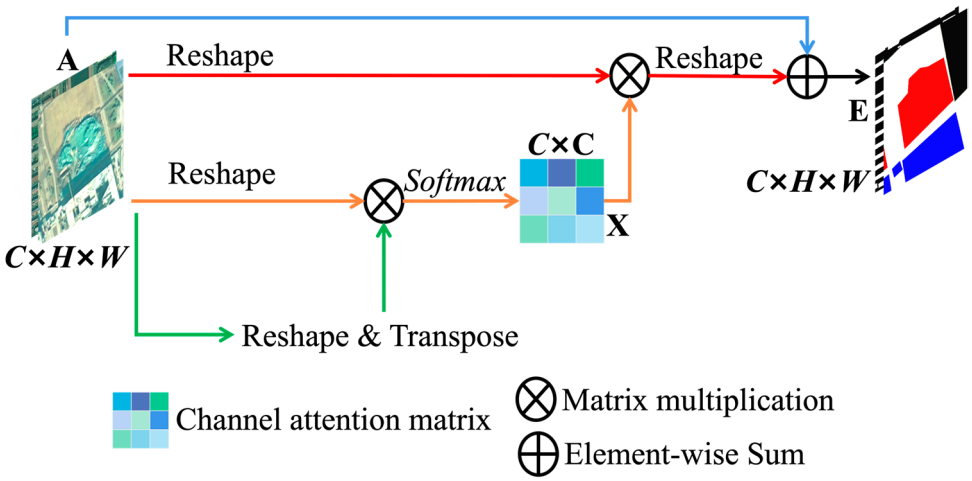


Figure 4. The detailed structure of channel attention.

The channel attention graph $\boldsymbol{X}$ is computed by $\boldsymbol{A}$. The computation of X_{ji} is shown below:

$$X_{ji} = \frac{exp\left[A_i \cdot \left(A' \right)_j^T \right]}{\sum_{i=1}^{C} exp\left[A'_i \cdot \left(A' \right)_j^T \right]}. \tag{3}$$

where X_{ji} represents the influence factor of the i-th channel on the j-th channel, A_i represents the value of the i-th element, and $\left(A' \right)_j^T$ represents the value of the j-th element of the transpose matrix.

The channel attention matrix is multiplied with the input feature map by a learnable parameter β, where β is initially zero. After that, it is summed element-wise with the input feature map, and the final output, ***E***, is calculated as follows:

$$E_j = \beta \sum_{i=1}^{C} \left(X_{ij} \cdot A_i \right) + A_j. \quad (4)$$

From formula (4), the final feature of each channel is the weighted sum of all channel features and the original channel features. Therefore, the channel attention module can make the feature map more prominent based on the feature information of different channels, enabling the neural network to perform more efficiently.

2.2.4. Dual Attention Mechanism Module

Challenges such as slow fitting speed, imprecise segmentation of edge targets, inconsistent segmentation of large-scale targets, and the presence of holes in the network's performance on remote sensing images are addressed by integrating both the positional attention module and the channel attention module. The DAMM (see the DAMM section in Figure 2) is introduced into the MT-AENet, where the upper branch represents the positional attention module, and the lower branch represents the channel attention module. The high-level feature maps output by the backbone network ResNet-101 are enhanced through these two branches and then element-wise summed up to obtain more refined high-level semantic information.

2.2.5. Integration of DAMM and ASPP Mechanisms

During the feature extraction process of the ResNet-101 backbone network, the overall receptive field grows gradually as the network layers progress. This leads to a loss of image information and a size reduction. To address this issue, the ASPP module is introduced to augment the network's receptive field without down-sampling, enhancing the network's ability to acquire multi-scale contextual information and mitigate information loss. To retain the excellent performance of the ASPP module while considering the outstanding feature enhancement capability of DAMM, a parallel structure that combines the DAMM structure with the ASPP module is designed and implemented. In this parallel structure, the backbone network extracts features from the image. The model is then divided into two branches, each handling the feature maps extracted by the backbone network. First, the feature map extracted from the backbone network is input into the DAMM in the upper branch. The lower branch uses the ASPP structure, and then, the feature maps processed by ASPP and DAMM are fused. Finally, the channel number of the fused feature map is reduced to 64 and input into the decoder, resulting in the final predicted segmentation map (see the DAMM and DS-ASPP sections in Figure 2).

2.2.6. Depthwise Separable-Atrous Spatial Pyramid Pooling Module

As deep learning network structures become more intricate, and with the introduction of the dual-attention mechanism, the network's parameters and computational load continuously increase, leading to more significant storage requirements and longer training times. To address this, the depthwise separable convolution (DSC) is introduced, which is a combination of depthwise convolution and pointwise convolution, into the ASPP structure, proposing the DS-ASPP module (see the DS-ASPP section in Figure 2). This structure not only directly reduces computations, thereby improving training speed, but also prevents the occurrence of gradient explosion, thereby increasing the predictive accuracy of the trained model.

In the process of convolution operation, if P_c is the number of DSC parameters, P_n is the number of standard convolution parameters, C_c is the amount of DSC computation, and C_n is the amount of standard convolution computation. The ratio of the number

of parameters of depthwise separable convolution and standard convolution is shown as follows:

$$\frac{P_c}{P_n} = \frac{C \cdot k^2 + C \cdot M}{C \cdot k^2 \cdot M} = \frac{1}{k^2} + \frac{1}{M}. \tag{5}$$

The ratio of the computational amount is shown as follows:

$$\frac{C_c}{C_n} = \frac{C \cdot k^2 \cdot H \cdot W + H \cdot W \cdot C \cdot M}{C \cdot k^2 \cdot M \cdot H \cdot W} = \frac{1}{k^2} + \frac{1}{M}. \tag{6}$$

where $\boldsymbol{C}$ is the number of input channels, $\boldsymbol{M}$ is the number of output channels, $\boldsymbol{k}$ is the size of the convolution kernel, and $H \times W$ is the size of output features.

As seen in formulas (5) and (6), the computational complexity of the depthwise separable convolution is significantly reduced compared to the standard convolution. Its computational amount is only about $1/k^2$ of the standard convolution, sacrificing only a minimal amount of precision in exchange for an order of magnitude reduction in computational amount.

2.3. Loss Function

This study employs the cross-entropy loss function to calculate the difference between the predicted category and the actual label for each pixel, serving as a crucial metric for optimizing model parameters.

2.3.1. Building Extraction Loss Function

The task of building extraction is a binary classification task (i.e., "Building" and "non-Building"). Therefore, a binary cross-entropy loss function is used. Employing the Sigmoid activation function as the output in the final layer, in the case of binary classification, where the model's predictions result in only two scenarios, the calculation formula is as follows:

$$\begin{gathered} L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot log p_i + (1 - y_i) \cdot log(1 - p_i)]. \\ \sigma(x) = \frac{1}{1 + exp(-x)}. \end{gathered} \tag{7}$$

where y_i represents the label of sample i with 1 for the positive class and 0 for the negative class, N is the number of pixels, p_i indicates the probability of sample i being predicted as the positive class, and $\sigma(x)$ is the Sigmoid activation function, ensuring predictions fall between 0 and 1.

2.3.2. Construction Waste Extraction Loss Function

In construction waste extraction, where the construction waste disposal sites are divided into four areas, it becomes a multi-label, multi-classification task. The cross-entropy loss function is employed, utilizing the SoftMax function as the network's final layer output. The multi-class scenario is essentially an extension of binary classification:

$$\begin{gathered} L_{CEL} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[y_{ij} \cdot log p_{ij} + \left(1 - y_{ij}\right) \cdot log\left(1 - p_{ij}\right)\right]. \\ \varphi(z)_i = \frac{exp(z_i)}{\sum_{j=1}^{M} \exp\left(z_j\right)}. \end{gathered} \tag{8}$$

where M represents the number of categories; N is the number of pixels; y_{ij} signifies the true category of sample j, taking 1 if the predicted category is the same and is otherwise 0; p_{ij} denotes the predicted probability of sample i belonging to category j; $\varphi(z)_i$ is the SoftMax function; and z_i is the output value of the i-th pixel.

2.3.3. MT-AENet Composite Loss Function

As the encoder produces feature maps with rich global information, combining the encoder features with decoder features helps retain more helpful information during the decoding stage, aiding segmentation accuracy. Therefore, this study calculates the loss function together for both the encoder and decoder, performing pixel-level predictions. The composite loss function is computed as follows:

$$Loss_{MT-AENet} = \alpha_1 \cdot L_{decoder} + \alpha_2 \cdot L_{encoder}. \quad (9)$$

where $Loss_{MT-AENet}$ is the composite loss function used by the MT-AENet, $L_{decoder}$ is the loss value from the decoder output, $L_{encoder}$ is the loss value from the encoder output, α_1 and α_2 are bias parameters, and $\alpha_1 + \alpha_2 = 1$. α_1, α_2 are used to adjust the weight of the loss value in the encoder and decoder sections.

2.4. Model for Calculating Annual Production of Construction Waste

This section provides a detailed description of the model used to calculate the annual production of construction waste. The following types of calculation formulas and parameter indicators are from the urban construction industry standard CJJ/T134-2019 [23] issued by the Ministry of Housing and Urban-Rural Development (MOHURD) of China, and it can be seen from the content of the standard that construction waste includes engineering waste and demolition waste. The detailed definitions are as follows:

- Engineering waste refers to discarded materials generated while constructing various buildings, structures, etc.
- Demolition waste refers to discarded materials generated during the demolition of various types of buildings, structures, etc.
- All types of construction waste, except the portion to be reused for resource utilization, will be put into designated construction waste disposal sites for temporary storage.

2.4.1. Generation Process of Engineering Waste and Demolition Waste

The production of engineering waste is derived from the area of newly constructed buildings in an urban region, and demolition waste production is obtained from demolished buildings. The specific calculation formulas are as follows:

$$M_g = R_g \cdot m_g. \quad (10)$$

where M_g represents the production of engineering waste in the urban region (t), R_g stands for the area of newly constructed buildings in the urban region (m^2), and m_g implies the basic production rate of engineering waste per unit area (t/m^2), ranging from 0.03 t/m^2 to 0.08 t/m^2.

$$M_c = R_c \cdot m_c. \quad (11)$$

where M_c represents the production of demolition waste in the urban region (t), R_c stands for the area of buildings being demolished in the urban region (m^2), and m_c implies the basic production rate of demolition waste per unit area (t/m^2), ranging from 0.8 t/m^2 to 1.3 t/m^2.

2.4.2. Calculation of Annual Production of Demolition Waste and Engineering Waste

From formulas (10) and (11), it is evident that when calculating either demolition waste or construction waste, obtaining the change in the area of urban or regional buildings is essential. The MT-AENet semantic segmentation network in this study possesses efficient multi-scale feature extraction capabilities, ensuring the high-precision extraction of buildings. The constructed building extraction model can identify and extract buildings across the entire region, revealing changes in regional buildings based on the data from different periods. To enhance the accuracy of the results, the study area is subdivided into smaller administrative regions for analysis.

To obtain the change in building area, the MT-AENet semantic segmentation network is used to extract the outlines of buildings. Given the spatial resolution of remote sensing satellites, the area represented by each pixel is known. The segmented images of the study area, organized by administrative areas, are fed into the network model to generate a predicted image for each specific region. Conducting a precise count of all white areas (i.e., buildings) in the predicted images enables calculating the total building area. The formula is as follows:

$$S_T = \sum_{i}^{T} SR \cdot SR \cdot p_i. \tag{12}$$

where S_T represents the total area of buildings in each administrative area (m^2), SR stands for the spatial resolution of remote sensing satellite, p_i implies the number of pixels in the white area of the predicted image of each administrative area, and T stands for the number of administrative regions.

The changes in building area are computed by analyzing remote sensing images from two consecutive years.

$$\Delta_S = S_{T2} - S_{T1}. \tag{13}$$

$$\begin{cases} M_g = R_g \cdot m_g;\ R_g = \Delta_S; \Delta_S > 0 \\ M_c = R_c \cdot m_c;\ R_c = \Delta_S; \Delta_S < 0 \end{cases}. \tag{14}$$

where S_{T2}, S_{T1} represent the building area at different times in the same region (m^2) and Δ_S stands for the change in building area (m^2).

2.5. Calculation of Volume of Construction Waste Landfill

A construction waste segmentation model is developed with extensive training, validation, and testing to accurately assess the study area's ability to handle construction waste. The model calculates weights based on the volume and average density of the pile and can quickly and accurately determine changes in construction waste within the landfill.

The MT-AENet semantic segmentation network is used to extract the contours of waste dump bodies. Subsequently, the volumes of the waste are estimated using the area of the constructed waste dump bodies multiplied by the average height of the dump bodies:

$$v_j = \sum_{i=1}^{N} SR \cdot SR \cdot q_i \cdot H. \tag{15}$$

where v_j represents the construction waste dump body volume per image (m^3), SR denotes the spatial resolution of the remote sensing image, q_i signifies the number of pixels in the red areas of each construction waste disposal site's predicted image, H denotes the waste dump bodies' average height (m), and N represents the number of construction waste disposal sites.

In formula (15), the average height H of the waste dump bodies is calculated based on the landfill requirements specified in CJJ/T134-2019 [23]. The spatial structure of the construction waste dump bodies consists of 5 layers. Table 1 provides a detailed description of the prescribed heights for each layer, where $h_1 \sim h_5$ represent the corresponding heights of each layer and H represents the sum of $h_1 \sim h_5$. It can be observed from Table 1 that the heights of the waste dump body range between 3.05 m and 5.05 m. Considering the irregular shape and non-uniform height of each pile area inside the landfill, the height of the pile is taken as the average value, i.e., H $\approx$ 4.0 m, for the convenience of the subsequent calculations.

Using two consecutive years of images from the construction waste disposal sites, the volume change in construction waste is calculated as follows:

$$v_1 = \sum_{j}^{N} v_j. \tag{16}$$

$$v_2 = \sum_{j}^{N} v_j. \tag{17}$$

$$\Delta_v = |v_2 - v_1|. \tag{18}$$

where N represents the number of images of the construction waste disposal sites, v_1 and v_2, respectively, denote the total volume of waste dump bodies in the study area at different times (m^3), and Δ_v represents the volume change of the waste (m^3).

Table 1. Types of layers and corresponding heights for landfill construction.

	Type of Layering	Height (cm)
1	Vegetative Layer	$h_1 \approx 15$
2	Drainage Layer	$h_2 \approx 30$
3	Impermeable Layer	$h_3 \approx 30$
4	Support and Ventilation Layer	$h_4 \approx 30$
5	Waste Layer	$200 \leq h_5 < 400$

The formula for calculating the weight based on the geometric model of construction waste dump bodies is as follows:

$$Emp = \rho_e \cdot \Delta_v. \tag{19}$$

where Emp represents the amount of construction waste (t), ρ_e is the average density of the waste dump bodies (t/m^3), with ρ_e ranging from 1.6 t/m^3 to 2.0 t/m^3, and Δ_v represents the volume change in the waste dump bodies (m^3).

3. Data and Experimental Setup

3.1. Study Area and Data Sources

This study focused on Changping District in Beijing, China for its high architectural density and diverse building types. The study area spans from 115°50′17″ to 116°29′49″E and 40°2′18″ to 40°23′13″N, encompassing a total area of 1343.5 km^2. The study heavily relies on the GF-2 data, an optical Earth observation satellite developed by the China National Space Administration (CNSA). GF-2 has a 0.8-m panchromatic camera and a 3.2-m multispectral camera, providing sub-meter spatial resolution and delivering high-quality, clear image information. Consequently, it has gained widespread application in remote sensing research [50,51]. Image data from Changping District collected by the GF-2 satellite in 2019 and 2020 were used for this study. Figure 5 presents an overview of the study area. The original GF-2 images of the study area (as shown in Figure 5c) capture urban, rural, and mountainous scenes. Furthermore, the images depicting buildings and construction waste disposal sites in the study area (as shown in Figure 5d) provide reliable data support for accurately estimating the annual production of construction waste in subsequent analyses.

3.2. Data Preprocessing

This section initiates the preprocessing of the original GF-2 images. This study used ArcMap software (version number: 10.8) to obtain vector maps and remote sensing images of Changping District, Beijing. These data were geographically calibrated and incorporated spatial coordinate information. Subsequently, the calibrated images were segmented into 500 × 500 pixel dimensions, resulting in 14,774 images. Following manual screening, 1000 clear images of buildings were retained. Concurrently, leveraging the geographic coordinates of the construction waste disposal sites in Changping District, 228 non-overlapping images of 500 × 500 pixels were located and segmented.

Semantic segmentation mask labels for buildings and construction waste disposal sites were manually annotated to form the basis for model training and accuracy evaluation. The dataset is presented in Figure 6. The dataset is relatively small due to the limitations of the actual sample size. Data augmentation techniques were applied to the existing dataset,

encompassing adjustments such as darkening, brightening, Gaussian noise, mirroring, and random scaling. This process yielded a final dataset comprising 6000 images of 500 × 500 pixels for the study area, with 4800 images allocated for training and 1200 for validation. Additionally, a dataset of 1368 images of 500 × 500 pixels for the construction waste disposal sites was obtained, with 1094 images used for training and 273 images for validation.

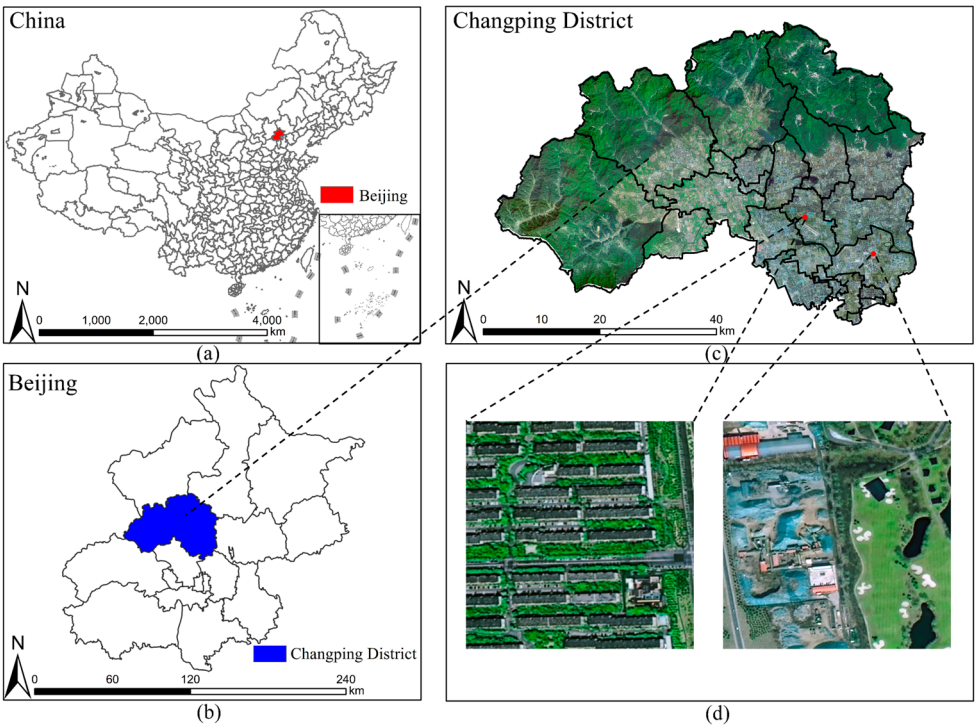


Figure 5. Overview of the study area: (**a**) Location of the study area in China, (**b**) Location of the study area in Beijing, (**c**) Original GF-2 images of the study area, and (**d**) Enlarged images of selected areas depicting buildings and construction waste disposal sites (highlighted with red dots).

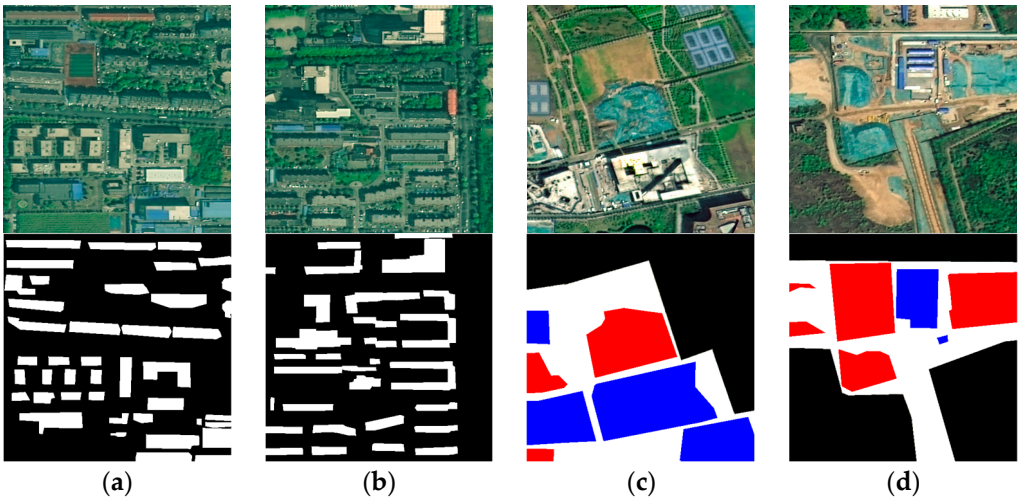


Figure 6. (**a**,**b**) depict images of buildings and their corresponding labeled images. The white area represents the building area, the black area indicates the non-building area. Similarly, (**c**,**d**) illustrate images of construction waste disposal sites and their corresponding labeled images. The white area represents vacant landfills, the black area indicates the image background, the blue area represents the engineering facility area, and the red area represents the dumping area.

3.3. Open-Source Construction Waste Disposal Site Datasets

The dataset was extended beyond Changping District to enhance data diversity and enhance the model's ability to generalize and identify construction waste within intricate environments. Specifically, images of construction waste disposal sites in Daxing District were annotated. These augmented data were seamlessly integrated, creating an enriched and openly accessible dataset called the Construction Waste Disposal Sites Dataset (CWDSD) [52]. Researchers can use this dataset to investigate the sources, types, distribution, and impact of construction waste. The CWDSD utilizes GF-2 satellite images with a spatial resolution of 80 cm GSD and Google Earth as the data source to construct a specific dataset of construction waste landfills in Changping and Daxing Districts of Beijing, China. The dataset comprises 3653 images of various construction waste disposal sites captured by satellites and detailed labeled images annotating different areas within the disposal sites. Researchers can employ the CWDSD dataset to train and evaluate the performance of semantic segmentation algorithms for detecting construction waste based on remote sensing images.

Additionally, it enables the quantitative analysis of solid waste production. The dataset is available at https://zenodo.org/record/8333888 (accessed on 13 November 2023). Table 2 lists several existing CNN-based waste datasets. It can be seen that solid waste datasets in aerial images can be used for detection, classification, or segmentation.

Table 2. Comparison of waste datasets for CNN.

	AerialWaste [47]	**SWAD [48]**	**CWDSD [52]**
Category	2	1	5
Instance	10,434	5562	10,959
Quantity	3478 + 6956	1597 + 399	2922 + 731
Distance	Medium and long	Long	Long
Task Type	Classification	Detection	Segmentation
Region	Lombardy, Italy	Henan, China	Beijing, China
Source	Airborne, WV-3, and Google Earth	WV-2 and SPOT	GF-2 and Google Earth

3.4. Comparative Methods

In this study, the MT-AENet is compared to classical networks in the semantic segmentation field, including UNet [53], SegNet [54], PSPNet [55], and DeepLabV3+ [56]. It is also compared to the latest published high-performance networks. For each dataset and model, the training epochs are fixed to 200, the batch size to 4, and the initial learning rate to 0.001 [57].

- The DSAT-Net [35] combines the strengths of CNNs and vision transformers in one architecture. It effectively reduces the computational costs and preserves spatial details during feature extraction.
- The ConvLSR-Net [42] introduces a novel, efficient long-short-range transformer (LSRFormer) to supplement global information. It addresses the need for local and global background information in aerial image segmentation.
- The SDSC-UNet [36] employs visual transformers and a U-shaped structure with multi-scale information capture, global dependency establishment, and shunted dual skip connection.

3.5. Experimental Setup

3.5.1. Experimental Environment

The compilation environment was Python 3.7, and the deep learning framework was PyTorch 1.9.0, using the PyCharm integrated development environment. The operating system was Windows 10 × 64, the CUDA version was 11.7, and the hardware platform consisted of an Intel(R) Core (TM) i9-10900X CPU@3.70 GHz (Intel Corporation is located

in Santa Clara, CA, USA), single CPU, 64.0 GB RAM, and Nvidia GeForce RTX 2080TI graphics card (Nvidia's headquarters is located in Santa Clara, CA, USA).

3.5.2. Evaluation Metrics

Five metrics are employed to provide a comprehensive evaluation, including Precision, Recall, Intersection over Union (IoU), F1 score (F1), and Bayesian Error Rate (BER). Recall and Precision are the most common evaluation metrics in semantic segmentation. Building extraction is a segmentation problem where building pixels are positives and non-building pixels are negatives. Since the focus is on changes in garbage heap areas, the pixels in that garbage disposal area are specified as positives, and pixels in other areas are negatives. Therefore, all predictions can be classified into a confusion matrix, including True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). All of these five metrics are defined (see (20)–(24)).

$$Precision = \frac{TP}{TP + FP}. \tag{20}$$

$$Recall = \frac{TP}{TP + FN}. \tag{21}$$

$$IoU = \frac{TP}{TP + FP + FN}. \tag{22}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} = \frac{2TP}{2TP + FN + FP}. \tag{23}$$

This study introduces the Bayesian error rate (BER) to assess the model's robustness, flexibility, and generalization in real-world applications. BER measures the extraction results' integrity, correctness, and quality and describes how the model can achieve minimum error boundary during extraction. This metric is crucial for multi-scale extraction, providing a fair evaluation of the deep model's performance in complex backgrounds [57].

$$BER = \frac{1}{2} \cdot \left(\frac{FN}{TP + FN} + \frac{FP}{FP + TN} \right). \tag{24}$$

4. Experimental Results

4.1. Building Extraction

4.1.1. Performance Evaluation

Table 3 presents the quantitative evaluation metrics for various models based on the Changping District, Beijing building dataset. It can be observed that the designed MT-AENet achieves the best results in all four metrics with the values of Precision (90.33%), Recall (87.90%), F1-Score (88.82%), and IoU (81.09%). The SegNet ranks second in Precision (88.39%), Recall (86.60%), F1 (87.15%), and IoU (78.64%). The UNet ranks third in Precision (86.87%), Recall (86.56%), F1 (86.37%), and IoU (76.94%). The MT-AENet is improved by 1.94% in Precision, 1.3% in Recall, 1.67% in F1, and 2.45% in IoU compared to the second-best in each metric.

Table 3. Quantitative comparison for the building dataset (%).

Model	Precision	Recall	F1	IoU
UNet [53]	86.87	86.56	86.37	76.94
SegNet [54]	88.39	86.60	87.15	78.64
PSPNet [55]	80.61	74.53	76.66	63.66
DeepLabV3+ [56]	83.78	86.22	84.57	74.26
DSAT-Net [35]	84.46	81.48	82.29	72.54
ConvLSR-Net [42]	85.80	82.54	83.41	74.38
SDSC-UNet [36]	77.87	74.67	75.44	61.78
MT-AENet	90.33	87.90	88.82	81.09

4.1.2. Qualitative Analysis

After performance evaluation, Figure 7 presents the visualization of the building extraction. In Figure 7a,b, the results showcase scenarios where small and dense buildings in complex scenes are obscured by surrounding large buildings, trees, roads, and other interfering objects, making building detection and segmentation more challenging. The UNet, PSPNet, and SDSC-UNet segmentation results suffer from boundary blurriness. At the same time, the proposed model is less affected by these interferences, exhibiting a more precise identification of building contours with fewer introduced errors. Figure 7c,d illustrates examples of extracting large buildings in urban scenes. When dealing with large buildings, a fixed receptive field may lead to discontinuous holes, failing to capture complete information about large structures and resulting in fragmented and blurred identified buildings. The existing network structure makes it difficult to solve this problem effectively. In contrast, the MT-AENet, utilizing depthwise separable dilated convolutions to expand the receptive field, demonstrates improved connectivity, effectively addressing such challenges.

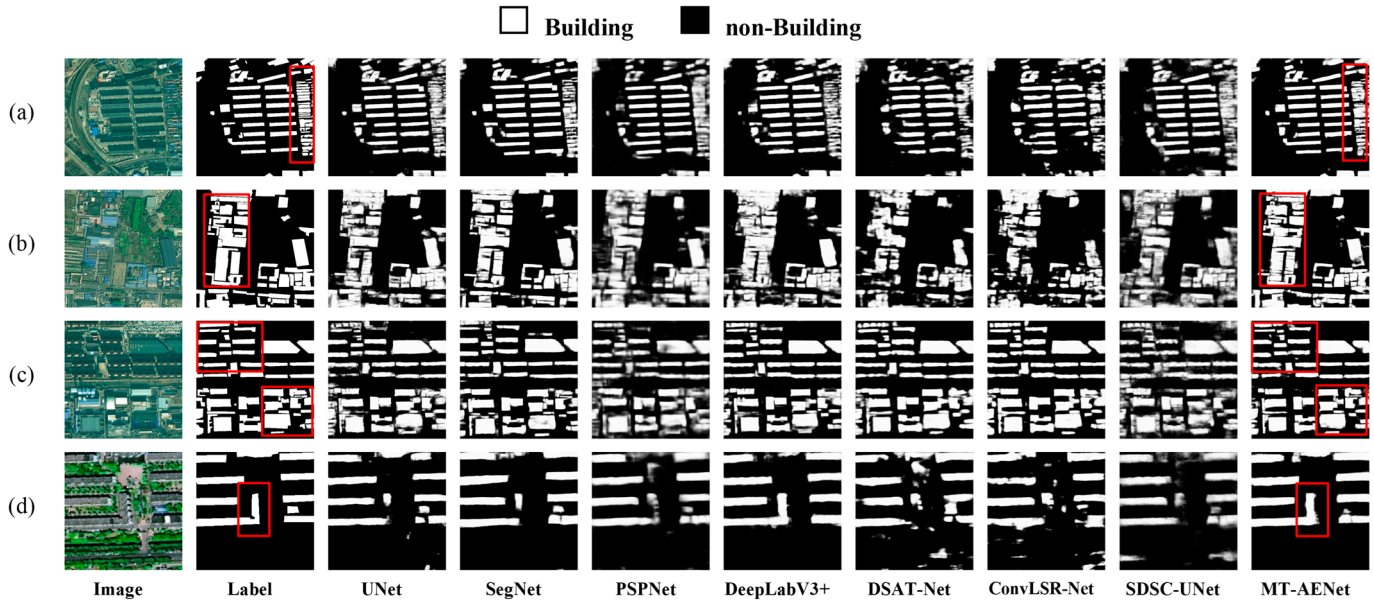


Figure 7. Results of different extraction methods on the building dataset in Changping District, Beijing. The original high-resolution image, mask-labeled image, UNet, SegNet, PSPNet, DeepLabV3+, DSAT-Net, ConvLSR-Net, SDSC-UNet, and MT-AENet prediction results are shown from left to right. (**a**,**b**) represent examples of extracting small dense buildings in complex scenes. (**c**,**d**) illustrates examples of extracting large buildings in urban scenes.

4.2. Construction Waste Extraction

4.2.1. Performance Evaluation

Table 4 presents the quantitative evaluation metrics for different models based on the CWDSD dataset. It is evident that the proposed MT-AENet, leveraging attention-enhancement and feature fusion mechanisms, achieves the highest scores on all four metrics with the values of Precision (90.43%), Recall (90.96%), F1-Score (89.40%), and IoU (83.35%). This is attributed to its preservation of more shallow-layer information in the network and the fusion of low-level features with high-level features, thereby enhancing the accuracy of segmentation boundaries. The UNet secures the second-best performance in Recall (89.92%) and IoU (82.54%). The SegNet ranks second in Precision (88.86%) and F1 (88.28%). Compared to the underperforming PSPNet, SDSC-UNet, and DeepLabV3+, the improvement of the proposed network on the IoU metric reaches up to 21%, and improvements on other metrics are consistently over 10%. The MT-AENet is improved

by 1.57% in Precision, 1.04% in Recall, 1.12% in F1, and 0.81% in IoU compared to the second-best in each metric.

Table 4. Quantitative comparison for the CWDSD dataset (%).

Model	Precision	Recall	F1	IoU
UNet [53]	87.67	89.92	88.16	82.54
SegNet [54]	88.86	88.70	88.28	82.45
PSPNet [55]	79.05	76.27	75.85	65.32
DeepLabV3+ [56]	77.85	73.45	72.37	62.35
DSAT-Net [35]	83.31	84.33	82.95	74.93
ConvLSR-Net [42]	85.45	86.38	85.10	78.40
SDSC-UNet [36]	76.27	77.07	75.38	64.02
MT-AENet	90.43	90.96	89.40	83.35

4.2.2. Qualitative Analysis

After performance evaluation, Figure 8 presents the visualization of the construction waste extraction. The comparative results indicate that some low-level features are excessively used due to the lack of attention modules. This leads to over-segmentation and fragmented construction waste extraction results from the existing models. Figure 8a illustrates an example of a waste pile area affected by building shadows, where it can be observed that DeepLabV3+, the DSAT-Net, and the ConvLSR-Net exhibit poorer extraction results under the influence of shadows, while the MT-AENet effectively addresses such scenarios. Figure 8b showcases an example of distinguishing between engineering facilities and waste pile features, where all networks can effectively differentiate between different features, with the MT-AENet excelling in contour segmentation details. Figure 8c displays an example where roads segment the waste pile. Figure 8d shows an example of extracting a waste pile under conditions of feature blurring. The DSAT-Net, ConvLSR-Net, and SDSC-UNet poorly recognize large waste piles, identifying most landfills as background areas. The UNet and SegNet have very similar results. The PSPNet and DeepLabV3+ misidentify the landfill area as open space, while the MT-AENet performs closer to the ground truth.

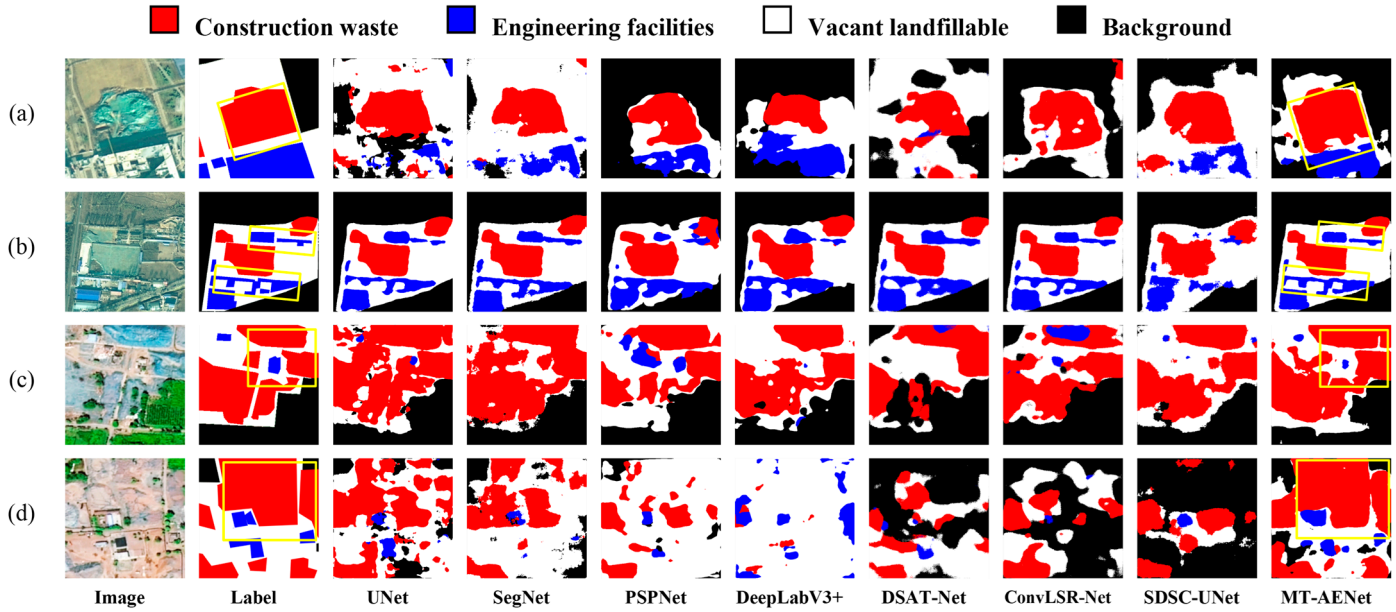


Figure 8. Results of different extraction methods on the CWDSD. The original high-resolution image, mask-labeled image, UNet, SegNet, PSPNet, DeepLabV3+, DSAT-Net, ConvLSR-Net, SDSC-UNet, and MT-AENet prediction results are shown from left to right. (**a**) illustrates an example of a waste pile area affected by building shadows, (**b**) showcases an example of distinguishing between engineering

facilities and waste pile features, (**c**) displays an example where roads segment the waste pile. (**d**) shows an example of extracting a waste pile under conditions of feature blurring.

4.3. BER Evaluation

In Figure 9, quantitative results are presented by comparing various methods in terms of BER on the building dataset in Changping District, Beijing and the CWDSD dataset. Figure 9 shows that the MT-AENet has the lowest error rates for the Changping District building dataset (6.78%) and the CWDSD dataset (5.46%). The SegNet achieves the second-lowest error rate for the Changping District building dataset (7.57%), indicating a 0.79% increase compared to the MT-AENet. The UNet ranks second-lowest regarding the CWDSD dataset, with an error rate of 6.16%, representing a 0.7% increase compared to the MT-AENet. These results suggest that the MT-AENet has a lower Bayesian error rate.

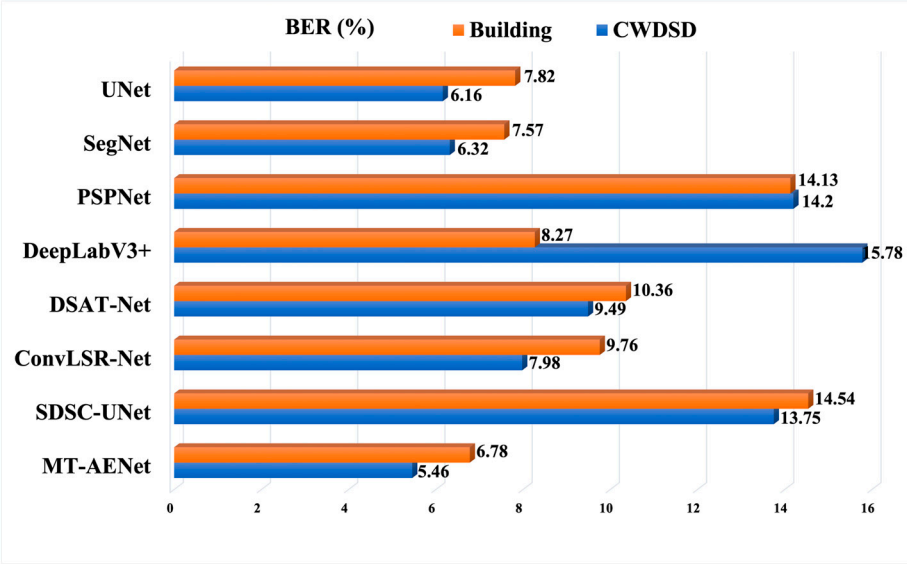


Figure 9. Comparison of BER for the building and CWDSD datasets.

On the one hand, this is attributed to introducing an attention-enhancement mechanism in the encoder, reducing noise interference. On the other hand, adopting ResNet-101 as a deeper backbone network enables better capture of complex features and relationships in the data, consequently reducing the error rate. The lower BER indicates that the MT-AENet is more robust and versatile in handling complex scenarios in authentic tasks.

4.4. Ablation Study

4.4.1. Ablation Experiments on the CWDSD Dataset

This section systematically analyzes the impact of hyperparameters and structural modifications on the MT-AENet. The ablation experiments use the controlled variable method to showcase the efficacy of incorporating the feature fusion module, DS-ASPP module, and DAMM in serial or parallel. Precision, Recall, F1-Score, and Model Size are key evaluation metrics.

Table 5 presents the quantitative analysis results of the ablation experiments on the CWDSD dataset, organized into five groups through the controlled variable method. The first group, as a baseline, exhibits the lowest performance in Precision (88.26%), Recall (89.06%), and F1 (88.42%). Comparatively, the second group, incorporating feature fusion, showcases improvements of 0.20%, 0.22%, and 0.30% in Precision, Recall, and F1, respectively. This improvement is attributed to addressing the deficiency in multi-scale feature fusion by incorporating mid-level features of the original image. In contrast to the second group, the third group, introducing DAMM serially with ASPP, shows an increase

of 0.57% in Precision, 1.01% in Recall, and 0.18% in F1. In the fourth group, where DAMM is parallel to ASPP, Precision is improved by 1.4%, Recall by 0.67%, and F1 by 0.5% compared to the serial structure of the third group. It is evident that both serial and parallel configurations of DAMM and ASPP effectively enhance network performance, with the parallel configuration demonstrating superior performance, particularly in addressing the deficiencies of the original network. However, the introduction of DAMM increases the model's parameter count, resulting in a model size of 203 M, an increase of 25 M compared to the network that is improved solely with ResNet-101. The fifth group introduces deep separable dilated convolutions to preserve high-precision extraction while minimizing model size and reducing training time. The results indicate a marginal decrease in Precision by 0.06%, Recall by 0.62%, and F1 by 0.44%. However, this minor precision loss is deemed acceptable for a reduction in model size by 13 M. In deep learning, numerous methods enhance model accuracy but often come with a significant increase in model parameters. Researchers must minimize model size to ensure better embedding and real-time capability while lowering training costs.

Table 5. Ablation experiment results on the CWDSD (%).

ResNet-101	FF	DAMM-S	DAMM-P	DS-ASPP	Precision	Recall	F1	Size (MB)
√					88.26	89.06	88.42	178
√	√				88.46	89.28	88.72	179
√	√	√			89.03	90.29	88.90	199
√	√		√		90.43	90.96	89.40	203
√	√		√	√	90.37	90.34	88.96	190

FF stands for Feature Fusion, DAMM-S stands for DAMM in series with ASPP, and DAMM-P stands for DAMM in parallel with ASPP.

4.4.2. Ablation Experiments on the Building Dataset

Table 6 presents the quantitative results of the ablation experiments on the building dataset in Changping District, Beijing. The experimental outcomes are generally consistent with those obtained for the CWDSD dataset. The first group ranks lowest in Precision (82.45%), Recall (84.27%), and F1 (83.25%). In contrast, compared to the first group, the fourth experimental group shows an improvement in Precision by 7.88%, Recall by 3.63%, and F1 by 5.57%. Simultaneously, the model's parameter performance aligns with the conclusions drawn for the CWDSD dataset.

Table 6. Ablation experiment results for the building dataset (%).

ResNet-101	FF	DAMM-S	DAMM-P	DS-ASPP	Precision	Recall	F1	Size (MB)
√					82.45	84.27	83.25	178
√	√				86.87	84.83	85.45	179
√	√	√			87.06	85.53	86.00	199
√	√		√		90.33	87.90	88.82	203
√	√		√	√	87.20	85.63	86.12	190

FF stands for Feature Fusion, DAMM-S stands for DAMM in series with ASPP, and DAMM-P stands for DAMM in parallel with ASPP.

4.4.3. Model Parameter Settings

There are other parameter settings in the modeling that affect the extraction results. Table 7 reports the effect of the composite loss function weight settings of the encoder and decoder on the results. For simplicity, all experiments in this section were conducted on the CWDSD dataset. In experiment 1, the loss values of the high-level feature maps generated by the encoder were assigned a larger weight. In experiment 2, equal weights were given. In experiment 3, the cross-entropy loss between the predicted map by the decoder and the label map was assigned a larger weight. As shown in Table 7, experiment 3 ranks highest in Precision (90.43%), F1 (89.40%), and IoU (83.35%) indicators. The ablation studies

indicate that increasing the weight of the decoder loss in the MT-AENet has a noticeable improvement in Precision, F1, and IoU metrics, but it also has a subtle impact on Recall.

Table 7. Comparison of different weight settings between encoder and decoder loss functions (%).

ID	Loss	Precision	Recall	F1	IoU
1	$0.3L_{decoder} + 0.7L_{encoder}$	88.32	90.39	88.74	82.37
2	$0.5L_{decoder} + 0.5L_{encoder}$	85.92	91.94	88.16	81.41
3	$0.7L_{decoder} + 0.3L_{encoder}$	90.43	90.96	89.40	83.35

4.5. Computational Complexity Study

To validate the flexibility and universality of the proposed MT-AENet, this section conducts a comprehensive computational complexity study on several networks, focusing on training time, speed, floating-point operations (FLOPs), and the number of parameters. Training time is the average time required to train one round using the same dataset. Speed is the number of images processed per second during training. FLOPs are calculated based on an input size of $1 \times 3 \times 512 \times 512$. All experiments were performed on the same device.

Table 8 shows that the FLOPs for PSPNet are about three times that of the MT-AENet. Compared to all the comparison models, the MT-AENet has the fastest training speed. This fully demonstrates the efficiency and flexibility of the MT-AENet, which can adapt to different types and scales of buildings or construction waste, which is crucial for complex real-world scenarios. However, the performance of the MT-AENet in terms of the number of parameters needs improvement compared to simpler network structures like the UNet, SegNet, and SDSC-UNet. Selecting a lightweight model to achieve optimal extraction in a shorter time is a challenge that will be addressed in future research.

Table 8. Complexity assessment results for different methods.

Model	Training Time (s)	Speed (n/s)	FLOPs (G)	Parameters (M)
UNet [53]	170.432	8.027	124.527	13.396
SegNet [54]	160.204	8.539	160.977	29.445
PSPNet [55]	801.980	1.706	201.851	52.508
DeepLabV3+ [56]	182.739	7.486	83.429	54.709
DSAT-Net [35]	194.040	7.050	57.675	48.371
ConvLSR-Net [42]	219.071	6.245	70.832	68.037
SDSC-UNet [36]	199.625	6.853	21.546	21.320
MT-AENet	151.222	9.046	68.720	52.929

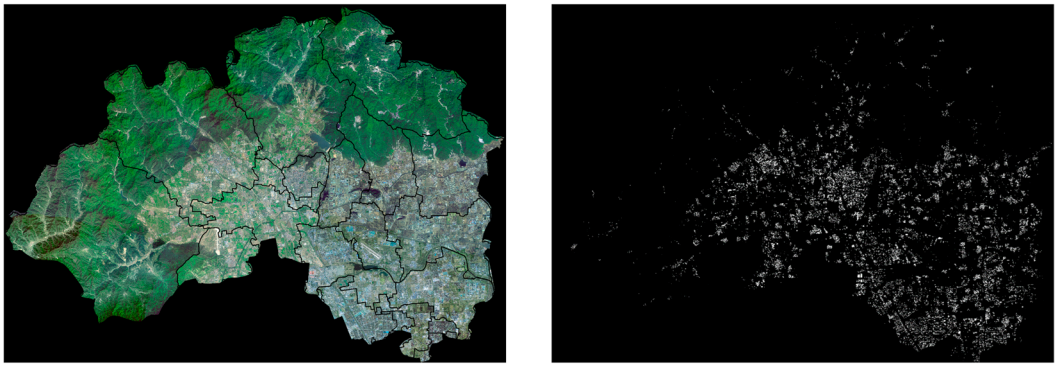
4.6. Annual Production Estimation of Construction Waste

In this section, the high-precision building segmentation model is employed to extract buildings across the entire Changping District, Beijing, facilitating the calculation of the annual production of construction waste. The extraction results are depicted in Figure 10, where the white areas represent the recognized building regions.

Figure 11 illustrates the area and variations of buildings in each administrative area segmented into 20 regions based on the administrative boundaries of Beijing. Using formulas (12), (13), and (14), where m_c ranges from 0.8 t/m^2 to 1.3 t/m^2 and m_g ranges from 0.03 t/m^2 to 0.08 t/m^2, the annual production of construction waste in each region is computed.

Table 9 illustrates the variations in building areas across different administrative regions of Changping District and the corresponding production of construction waste. According to the data presented in Table 9, the total building area in Changping District was 10,389.570 hm^2, or approximately 103.90 km^2, in 2019, and in 2020, it amounted to 10,086.387 hm^2, or approximately 100.86 km^2. The building area derived from the GIS platform for Changping District is 102.55 km^2. For a unit of 1 km^2, the error calculated in building area by this study is only 0.013 km^2, providing a relatively precise basis for

calculating the annual production of construction waste in Changping District, Beijing. The experimental data show that from 2019 to 2020, a cumulative total of 303.182 hm^2, or approximately 3.03 km^2, of buildings were demolished and renovated. The engineering waste generated during the urban renewal ranges from 24,916 tons to 66,443 tons, with an average of approximately 45,679.5 tons. The demolition waste generated is estimated from 3,089,887 tons to 5,021,067 tons, with an average of approximately 4,055,477 tons. The final calculation estimates a total of 4,101,156.5 tons of construction waste generated.



(**a**) (**b**)

Figure 10. Building extraction results in Changping District, Beijing: (**a**) Original GF-2 remote sensing image of the study area and (**b**) Distribution of identified buildings, with white areas indicating the recognized buildings and black areas representing the background.

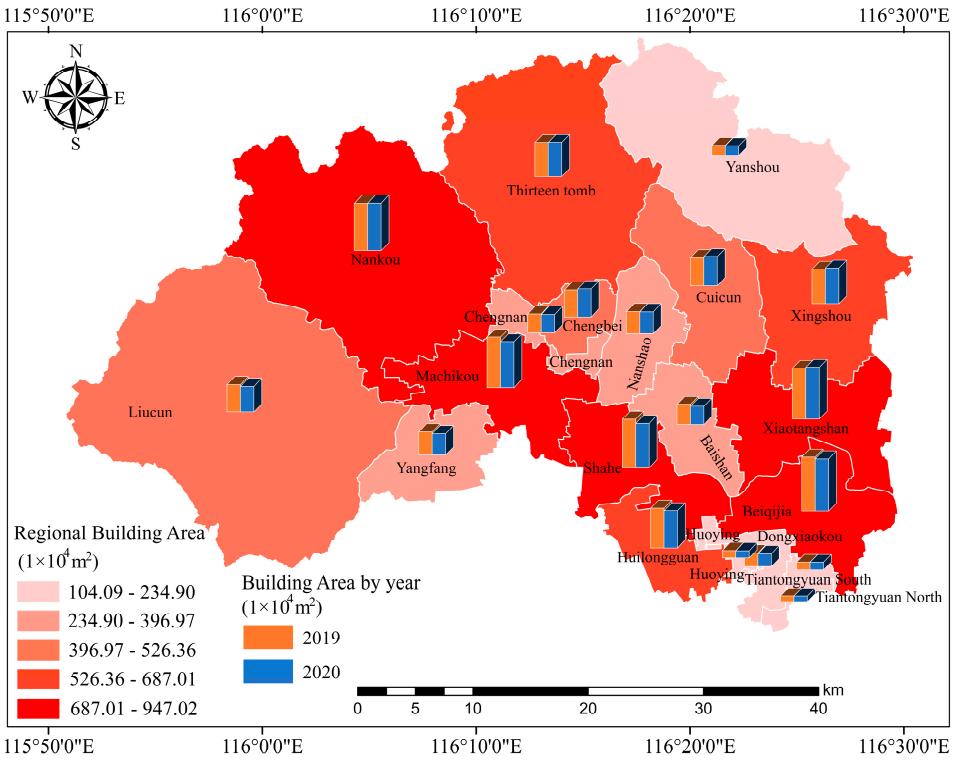


Figure 11. Change in the building area of each administrative area in Changping District.

The results suggest that from 2019 to 2020, Changping District predominantly engaged in the demolition of buildings, propelled by urban renewal initiatives in line with Beijing's recent focus on waste reduction. According to the "2023 Beijing Government Work Report," over the past five years, Beijing has demolished 240 million m^2 of illegally constructed buildings [58]. In 2020 alone, the estimated total construction waste generation reached

52,000,000 tons. Given that Changping District constitutes approximately 8% of Beijing's total area, a rough estimate indicates the annual demolition of buildings of an average of 3.84 km^2, resulting in 4,160,000 tons of construction waste. This implies an impressive accuracy rate of 98.59% in the calculations presented in this study, further validating the credibility of the experimental results.

Table 9. Changes in building area and construction waste production in Changping district, Beijing, China.

Administrative Area	2019	2020	Change (10^4 m^2)	Engineering Waste		Demolition Waste	
	Building Area (10^4 m^2)			At Least (t)	At Most (t)	At Least (t)	At Most (t)
Baishan Town	369.644	333.440	−36.204			289,631.744	470,651.584
Beiqijia Town	976.173	942.694	−33.479			267,835.904	435,233.344
Chengbei Street	494.522	514.295	19.773	5931.878	15,818.342		
Chengnan Street	341.938	330.765	−11.174			89,383.424	145,248.064
Cuicun Town	508.425	526.358	17.933	5379.917	14,346.445		
Dongxiaokou District	247.839	234.896	−12.943			103,542.272	168,256.192
Huilongguan District	734.752	687.013	−47.739			381,913.600	620,609.600
Huoying Street	130.196	126.665	−3.531			28,248.064	45,903.104
Liucun Town	496.992	463.736	−33.257			266,053.632	432,337.152
Machikou District	918.954	831.378	−87.576			700,609.536	1,138,490.496
Nankou District	854.672	860.037	5.365	1609.478	4291.942		
Nanshao Town	397.953	394.846	−3.107			24,855.040	40,389.440
Shahe District	875.982	801.448	−74.534			596,271.616	968,941.376
Thirteen tomb Town	614.671	617.267	2.596	778.906	2077.082		
Tiantongyuan North	144.077	142.330	−1.746			13,971.456	22,703.616
Tiantongyuan South	105.562	104.087	−1.475			11,797.504	19,170.944
Xiaotangshan Town	930.129	947.022	16.893	5067.878	13,514.342		
Xingshou Town	630.362	650.856	20.493	6147.994	16,394.650		
Yanshou Town	186.360	180.279	−6.081			48,649.728	79,055.808
Yangfang Town	430.365	396.975	−33.390			267,123.712	434,076.032
Changping District	10,389.570	10,086.387	−303.182	24,916.051	66,442.803	3,089,887.232	5,021,066.752

4.7. Annual Resource Conversion Rate of Construction Waste

The disposal of urban-generated construction waste typically follows two main pathways. Aside from construction waste undergoing resource recovery and reuse, the remaining waste is generally landfilled in construction waste disposal sites. To analyze the landfill volume and resource conversion of regional construction waste and determine the urban annual processing capacity for construction waste, a high-precision construction waste segmentation model is employed to extract all construction waste disposal sites in Changping District. The urban annual capacity to process construction waste can be determined by analyzing the relationship between actual construction waste production and the landfill volume.

According to formula (15), where SR = 0.8 m, the cumulative area of construction waste piles can be accurately calculated. Table 10 presents the extraction results for the area of each region in construction waste disposal sites, where red represents construction waste piles. Visualization analysis in Figure 12 illustrates the growth in the scale of construction waste disposal sites across the entire district from 2019 to 2020.

Table 10. Extraction results of the area of construction waste disposal sites in each region.

Area (m^2)	2019 (m^2)	2020 (m^2)	Change in Area (m^2)
Construction waste	8,199,598.08	8,512,355.84	312,757.76
Engineering facilities	898,778.24	908,949.76	10,171.52
Vacant landfills	5,490,789.12	6,335,324.80	844,535.68
Background	30,050,834.56	28,883,369.60	−1,167,464.96

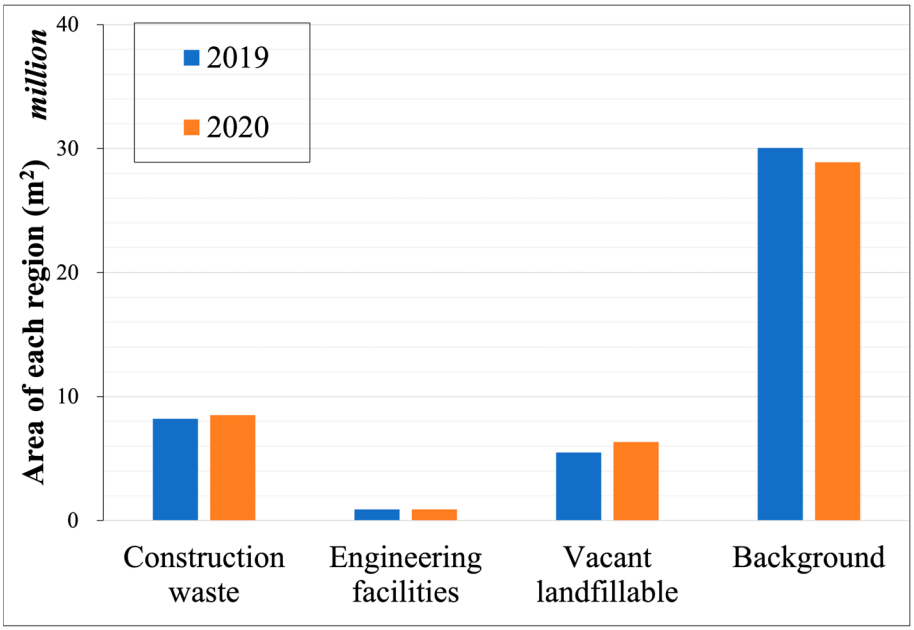


Figure 12. The area of each area within the construction waste disposal sites in Changping District, Beijing in 2019 and 2020.

Table 11 illustrates the relationship between the number of landfill waste pixels, area, volume, and weight. Using the formulas (15), (16), (17), and (18), where SR = 0.8 m and H ≈ 4.0 m, the variation in the volume of landfill waste can be computed. Combining this with formula (19), where $\rho_e \approx 1.8$ t/m^3, the landfill volume of construction waste can be accurately estimated. Table 11 shows that in the construction waste disposal sites in Changping District, Beijing, approximately 32,798,392.32 m^3 of construction waste was deposited in 2019, weighing about 59,037,106.176 tons. In 2020, approximately 34,049,423.36 m^3 of construction waste was deposited, weighing about 61,288,962.048 tons. The difference over the two years indicates an increase in construction waste in the disposal sites of Changping District by approximately 2,251,855.872 tons.

Table 11. Calculation of construction waste landfill volume.

Landfill Waste	2019	2020
Number of red pixels	12,811,872	13,300,556
Area (m^2)	8,199,598.08	8,512,355.84
Volume (m^3)	32,798,392.32	34,049,423.36
Weight (t)	59,037,106.176	61,288,962.048
Landfill volume (t)	2,251,855.872	

Table 12 analyzes the annual processing capacity of construction waste in Changping District, Beijing. Using the precise calculation of the change in building area over two consecutive years, the actual annual production of construction waste is approximately 4,101,156.500 tons. Since urban-generated construction waste mainly undergoes two pathways, i.e., landfilling and resource conversion, by extracting the increase in waste piles in the construction waste disposal sites, the landfill portion can be accurately calculated to be approximately 2,251,855.872 tons. Thus, the indirectly calculated resource conversion portion is approximately 1,849,300.628 tons, resulting in an annual resource conversion rate of 45.09%.

Table 12. Analysis of the annual processing capacity of construction waste in Changping district.

Type	Weight
Annual production	4,101,156.500 t
Landfill component	2,251,855.872 t
Reutilized component	1,849,300.628 t
Resource conversion rate	45.09%

5. Discussion

5.1. Difference between Existing Works and Our Approach

Traditional methods for estimating construction waste include the following: (1) Statistical data-based methods [59,60]: combining relevant statistical data to estimate the production of construction waste. This method heavily relies on historical data and statistical methods, which may result in biases due to data quality issues, thus lacking flexibility and accuracy in practical applications. (2) Construction engineering project list method [61]: analyzing the list of construction engineering projects, estimating construction waste generated by each project, and then summing them up to obtain the total construction waste. The accuracy of this method heavily depends on the accuracy and completeness of the list. Omissions or inaccuracies in the list will result in deviations in practical applications. (3) Building Information Modeling (BIM)-based estimation methods [62,63]: utilizing BIM technology to detail the modeling of buildings during the architectural design phase, including information such as materials, structures, and quantities, to estimate construction waste generation. However, the BIM technology requires professional software and technical support, as well as a large amount of architectural design and construction data, which are not easy to obtain or integrate, limiting the application scope of this method.

From this perspective, while there have been some advancements in construction waste estimation methods in recent years, the above methods share some common limitations. They heavily rely on prior data and are influenced by subjective factors, resulting in experimental results lacking verifiability. Additionally, the methods only apply to specific regions, lacking transferability, thus failing to meet the precise estimation of urban construction waste proposed in this study. HRSIs, on the other hand, exhibit strong timeliness, high accuracy, and macroscopic observations, making them one of the most effective methods in urban dynamic monitoring. They can analyze urban-level construction waste disposal capacity more accurately and efficiently. Furthermore, estimating construction waste production based on area change is also popular. For instance, Miatto et al. [17] employed a spatially explicit analysis method to estimate the amount of waste generated by demolition activities by analyzing maps and aerial images to calculate the lifespan of buildings. Wu et al. [64] proposed an off-site snapshot method for estimating construction waste, selecting three construction sites, and capturing images at their initial, mid-term, and final stages to quickly obtain reliable waste generation information and production rates. The amount of construction waste can be effectively calculated by analyzing the changes in construction waste in images from different periods in the same area.

The method proposed in this study is also based on area change for estimating construction waste production, but it differs from the research mentioned above. By contrasting it against state-of-the-art methods in the semantic segmentation field, a high-precision MT-AENet is proposed, effectively enhancing the capability of extracting features from HRSIs. Furthermore, by integrating multiple data sources, this study achieves the demand for the dynamic monitoring of regional buildings and tracking changes in construction waste disposal sites, thereby achieving a precise estimation of regional construction waste production. This study holds theoretical and practical significance in quantifying and managing construction waste: (1) The verification of existing conclusions has been carried out, providing a more systematic, scientific, and efficient processing method, which enhances operational efficiency. (2) In the current field of waste estimation, this study breaks away from the predicament of traditional manual estimation, making the method

applicable to different environments and cities, thus enhancing its practicality. (3) The computational results of this study also contribute to the current estimation conclusions, particularly in spatial refinement statistics, achieving collaborative conclusions among counties. (4) Time-saving and efficient construction waste calculation methods are more attractive and practical for governments and related enterprises. They can be directly used to guide the formulation of urban renewal policies and provide important data support for the resourceful handling of urban construction waste.

5.2. Limitations of the Methodology

In this study, we have proposed a novel method to accurately estimate the production and landfill volume of construction waste. The model used to extract features from HRSIs has achieved satisfactory accuracy. The computed results of construction waste also match the statistical data released by the government, demonstrating the method's effectiveness. However, we have also identified some existing issues. For example, in Section 2.4, the calculation of construction waste and demolition waste relies largely on prior parameters, such as the basic production rate of engineering waste per unit area (m_g), the basic production rate of demolition waste per unit area (m_c), and the average density of the waste dump bodies (ρ_e). Although these parameter indicators are from the "Urban Construction Industry Standard" CJJ/T134-2019 [23] issued by the MOHURD of China, possessing official authority, the accurate values of these parameters are indeed not easy to obtain as only a range of values can be acquired under general circumstances due to their susceptibility to various factors, including regional characteristics, building types, and construction methods. We further emphasize this point in the discussion to avoid any misunderstanding of the proposed method. The accuracy of the a priori parameters will be further improved in future work.

6. Conclusions and Future Directions

6.1. Conclusions

This study investigates the CNN-based semantic segmentation of HRSIs to identify and extract buildings and construction waste. Considering the limitations of traditional recognition network architectures, a novel encoder–decoder structure is designed, constructing the Multi-scale Target Attention-Enhanced Network by integrating multi-scale features. The MT-AENet leverages contextual information from HRSIs more effectively, enhancing the model's accuracy in recognition. Buildings in the study area are extracted through the MT-AENet. The engineering and demolition waste are calculated based on the increased and decreased building areas from HRSI data over two consecutive years. Consequently, the annual production of construction waste is calculated. Simultaneously, the MT-AENet extracts and analyzes the change in construction waste in the construction waste disposal sites during the same period. The annual production of construction waste and the annual resource conversion rate in the regional construction waste are accurately estimated. The experimental results indicate the following conclusions.

First, for the identification and extraction of buildings, the MT-AENet outperforms traditional networks with an improvement in Precision, Recall, F1, and IoU by 0.33%, 1.94%, 1.3%, 1.67%, and 2.45%, respectively. The BER is also reduced by 0.79%. For the identification and extraction of construction waste, the MT-AENet improves in Precision, Recall, F1, and IoU by 1.85%, 1.57%, 1.04%, 1.12%, and 0.81%, respectively. The BER is reduced by 0.7%. The MT-AENet is a high-precision and flexible model for dynamic recognition through HRSIs.

Second, by dividing the study area into smaller administrative regions, changes in buildings can be monitored dynamically, and the increase or decrease in the area of buildings in each region can be analyzed quickly and accurately. The calculations reveal that from 2019 to 2020, approximately 3.03 km^2 of buildings were dismantled and renovated in Changping District. The engineering waste generated during urban renewal ranged from 24,916 tons to 66,443 tons, with an average of approximately 45,679.5 tons. The demolition

waste generated ranged from 3,089,887 tons to 5,021,067 tons, averaging approximately 4,055,477 tons. The estimated annual construction waste production is determined to be approximately 4,101,156.5 tons.

Third, construction waste can be extracted from the disposal sites in Changping District. The construction waste landfill volume was approximately 32,798,392.32 m^3, weighing 59,037,106.176 tons in 2019. In 2020, approximately 34,049,423.36 m^3 of construction waste was deposited, weighing about 61,288,962.048 tons. The difference over the two years indicates an increase in construction waste in the disposal sites in Changping District by approximately 2,251,855.872 tons. In summary, the indirectly calculated construction waste for resource conversion is approximately 1,849,300.628 tons, with an annual resource conversion rate of 45.09%.

6.2. Future Directions

Currently, this study is limited to Changping District in Beijing. In the future, remote sensing images collected from different areas will be used to expand this study into a more extensive research area. This expansion aims to estimate the annual production of construction waste for an entire city, province, country, or even more extensive region, providing comprehensive and timely data support for urban renewal. This will aid in formulating scientifically sound urban renewal plans, reducing cost risks, and achieving the sustainable development goals of urban renewal.

Author Contributions: L.H. performed the data processing and technical validation, conducted relevant experiments, and wrote the manuscript. S.L. determined the topic, supervised the study, and revised the manuscript. X.L. designed the methodology, supervised the study, and revised the manuscript. S.W. revised the manuscript. G.C. provided GF-2 images and revised the manuscript. Q.M. revised the manuscript. Z.F. guided the innovation of urban data governance application scenarios and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the project "Research on the application and practice of data sinking empowerment of grass-roots social governance based on urban planning coordination" of Beijing Big Data Centre, which is funded by the Natural Science Foundation of Beijing Municipality (No.9232008).

Data Availability Statement: The Construction Waste Disposal Sites Dataset (CWDSD) has been released on the Zenodo repository https://zenodo.org/record/8333888 (accessed on 13 November 2023).

Acknowledgments: The authors would like to thank the Beijing Big Data Centre for providing the images and location data. The authors would also like to thank the constructive comments from the editors and anonymous reviewers.

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

References

1. Zheng, H.W.; Shen, G.Q.; Wang, H. A Review of Recent Studies on Sustainable Urban Renewal. *Habitat Int.* **2014**, *41*, 272–279. [CrossRef]
2. Liang, L.; Wang, Z.; Li, J. The Effect of Urbanization on Environmental Pollution in Rapidly Developing Urban Agglomerations. *J. Clean. Prod.* **2019**, *237*, 117649. [CrossRef]
3. Sun, M.; Wang, J.; He, K. Analysis on the Urban Land Resources Carrying Capacity during Urbanization—A Case Study of Chinese YRD. *Appl. Geogr.* **2020**, *116*, 102170. [CrossRef]
4. Yıldız, S.; Kıvrak, S.; Gültekin, A.B.; Arslan, G. Built Environment Design—Social Sustainability Relation in Urban Renewal. *Sustain. Cities Soc.* **2020**, *60*, 102173. [CrossRef]
5. Akhtar, A.; Sarmah, A.K. Construction and Demolition Waste Generation and Properties of Recycled Aggregate Concrete: A Global Perspective. *J. Clean. Prod.* **2018**, *186*, 262–281. [CrossRef]
6. Lin, Y.; Zhang, H. Intra-Year Urban Renewal in Metropolitan Cities Using Time Series SAR and Optical Data. In Proceedings of the IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 6288–6291.

7. Vieira, C.S.; Pereira, P.M. Use of Recycled Construction and Demolition Materials in Geotechnical Applications: A Review. *Resour. Conserv. Recycl.* **2015**, *103*, 192–204. [CrossRef]
8. Yu, B.; Wang, J.; Li, J.; Zhang, J.; Lai, Y.; Xu, X. Prediction of Large-Scale Demolition Waste Generation during Urban Renewal: A Hybrid Trilogy Method. *Waste Manag.* **2019**, *89*, 1–9. [CrossRef] [PubMed]
9. China Population: Urbanization Rate: Usual Residence | Economic Indicators | CEIC. Available online: https://www.ceicdata.com/en/china/population-urbanization-rate/cn-population-urbanization-rate-usual-residence (accessed on 1 December 2023).
10. Liang, X.; Lin, S.; Bi, X.; Lu, E.; Li, Z. Chinese Construction Industry Energy Efficiency Analysis with Undesirable Carbon Emissions and Construction Waste Outputs. *Environ. Sci. Pollut. Res.* **2021**, *28*, 15838–15852. [CrossRef]
11. Zhu, J. China Construction and Demolition Waste Industry Market Report · EnvGuide. 2021. Available online: https://us.envguide.com/china-construction-and-demolition-waste-industry-market-report/ (accessed on 1 December 2023).
12. López Ruiz, L.A.; Roca Ramón, X.; Gassó Domingo, S. The Circular Economy in the Construction and Demolition Waste Sector—A Review and an Integrative Model Approach. *J. Clean. Prod.* **2020**, *248*, 119238. [CrossRef]
13. Wu, Z.; Yu, A.T.W.; Shen, L.; Liu, G. Quantifying Construction and Demolition Waste: An Analytical Review. *Waste Manag.* **2014**, *34*, 1683–1692. [CrossRef]
14. Hoang, N.H.; Ishigaki, T.; Kubota, R.; Tong, T.K.; Nguyen, T.T.; Nguyen, H.G.; Yamada, M.; Kawamoto, K. Waste Generation, Composition, and Handling in Building-Related Construction and Demolition in Hanoi, Vietnam. *Waste Manag.* **2020**, *117*, 32–41. [CrossRef]
15. Bakchan, A.; Faust, K.M. Construction Waste Generation Estimates of Institutional Building Projects: Leveraging Waste Hauling Tickets. *Waste Manag.* **2019**, *87*, 301–312. [CrossRef]
16. McBean, E.A.; Fortin, M.H.P. A Forecast Model of Refuse Tonnage with Recapture and Uncertainty Bounds. *Waste Manag. Res.* **1993**, *11*, 373–385. [CrossRef]
17. Miatto, A.; Schandl, H.; Forlin, L.; Ronzani, F.; Borin, P.; Giordano, A.; Tanikawa, H. A Spatial Analysis of Material Stock Accumulation and Demolition Waste Potential of Buildings: A Case Study of Padua. *Resour. Conserv. Recycl.* **2019**, *142*, 245–256. [CrossRef]
18. Llatas, C. A Model for Quantifying Construction Waste in Projects According to the European Waste List. *Waste Manag.* **2011**, *31*, 1261–1276. [CrossRef]
19. Guerra, B.C.; Bakchan, A.; Leite, F.; Faust, K.M. BIM-Based Automated Construction Waste Estimation Algorithms: The Case of Concrete and Drywall Waste Streams. *Waste Manag.* **2019**, *87*, 825–832. [CrossRef]
20. Lu, W.; Lou, J.; Webster, C.; Xue, F.; Bao, Z.; Chi, B. Estimating Construction Waste Generation in the Greater Bay Area, China Using Machine Learning. *Waste Manag.* **2021**, *134*, 78–88. [CrossRef]
21. Ramnarayan; Malla, P. A Machine Learning-Enhanced Method for Quantifying and Recycling Construction and Demolition Waste in India. In Proceedings of the 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), Raichur, India, 24–25 February 2023; pp. 1–7.
22. Bioucas-Dias, J.M.; Plaza, A.; Camps-Valls, G.; Scheunders, P.; Nasrabadi, N.; Chanussot, J. Hyperspectral Remote Sensing Data Analysis and Future Challenges. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–36. [CrossRef]
23. CJJT134-2019—PDF BOOK. Available online: https://www.mohurd.gov.cn/gongkai/zhengce/zhengcefilelib/201910/20191012_242186.html (accessed on 24 November 2023).
24. Domingo, N.; Batty, T. Construction Waste Modelling for Residential Construction Projects in New Zealand to Enhance Design Outcomes. *Waste Manag.* **2021**, *120*, 484–493. [CrossRef]
25. Hoang, N.H.; Ishigaki, T.; Kubota, R.; Tong, T.K.; Nguyen, T.T.; Nguyen, H.G.; Yamada, M.; Kawamoto, K. Financial and Economic Evaluation of Construction and Demolition Waste Recycling in Hanoi, Vietnam. *Waste Manag.* **2021**, *131*, 294–304. [CrossRef]
26. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a Convolutional Neural Network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6.
27. Lin, H.; Hao, M.; Luo, W.; Yu, H.; Zheng, N. BEARNet: A Novel Buildings Edge-Aware Refined Network for Building Extraction fromHigh-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6005305. [CrossRef]
28. Kang, W.; Xiang, Y.; Wang, F.; You, H. EU-Net: An Efficient Fully Convolutional Network for Building Extraction from Optical Remote Sensing Images. *Remote Sens.* **2019**, *11*, 2813. [CrossRef]
29. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A Fully Convolutional Neural Network for Automatic Building Extraction fromHigh-Resolution Remote Sensing Images. *Remote Sens.* **2020**, *12*, 1050. [CrossRef]
30. He, H.; Wang, S.; Zhao, Q.; Lv, Z.; Sun, D. Building Extraction Based on U-Net and Conditional Random Fields. In Proceedings of the 2021 6th International Conference on Image, Vision and Computing (ICIVC), Qingdao, China, 23–25 July 2021; pp. 273–277.
31. Chen, M.; Wu, J.; Liu, L.; Zhao, W.; Tian, F.; Shen, Q.; Zhao, B.; Du, R. DR-Net: An Improved Network for Building Extraction from High Resolution Remote Sensing Image. *Remote Sens.* **2021**, *13*, 294. [CrossRef]
32. Wang, H.; Miao, F. Building Extraction from Remote Sensing Images Using Deep Residual U-Net. *Eur. J. Remote Sens.* **2022**, *55*, 71–85. [CrossRef]
33. Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Plaza, A.; Chanussot, J. Graph Convolutional Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5966–5978. [CrossRef]
34. Wang, A.; Zhang, P. Automatic Building Extraction Based on Boundary Detection Network in Satellite Images. In Proceedings of the 2022 29th International Conference on Geoinformatics, Beijing, China, 15–18 August 2022; pp. 1–7.

35. Zhang, R.; Wan, Z.; Zhang, Q.; Zhang, G. DSAT-Net: Dual Spatial Attention Transformer for Building Extraction fromAerial Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6008405. [CrossRef]
36. Zhang, R.; Zhang, Q.; Zhang, G. SDSC-UNet: Dual Skip Connection ViT-Based U-Shaped Model for Building Extraction. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6005005. [CrossRef]
37. Chen, Q.; Cheng, Q.; Wang, J.; Du, M.; Zhou, L.; Liu, Y. Identification and Evaluation of Urban Construction Waste with VHR Remote Sensing Using Multi-Feature Analysis and a Hierarchical Segmentation Method. *Remote Sens.* **2021**, *13*, 158. [CrossRef]
38. Davis, P.; Aziz, F.; Newaz, M.T.; Sher, W.; Simon, L. The Classification of Construction Waste Material Using a Deep Convolutional Neural Network. *Autom. Constr.* **2021**, *122*, 103481. [CrossRef]
39. Sun, X.; Liu, Y.; Yan, Z.; Wang, P.; Diao, W.; Fu, K. SRAF-Net: Shape Robust Anchor-Free Network for Garbage Dumps in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6154–6168. [CrossRef]
40. Li, H.; Hu, C.; Zhong, X.; Zeng, C.; Shen, H. Solid Waste Detection in Cities Using Remote Sensing Imagery Based on a Location-Guided Key Point Network with Multiple Enhancements. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 191–201. [CrossRef]
41. Liu, X.; Liu, Y.; Du, M.; Zhang, M.; Jia, J.; Yang, H. Research on Construction and Demolition Waste Stacking Point Identification Based on DeeplabV3+. *Bull. Surv. Mapp.* **2022**, 16–19+43. [CrossRef]
42. Zhang, R.; Zhang, Q.; Zhang, G. LSRFormer: Efficient Transformer Supply Convolutional Neural Networks with Global Information for Aerial Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5610713. [CrossRef]
43. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
44. Zhou, Y.; Chen, Z.; Wang, B.; Li, S.; Liu, H.; Xu, D.; Ma, C. BOMSC-Net: Boundary Optimization and Multi-Scale Context Awareness Based Building Extraction fromHigh-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5618617. [CrossRef]
45. Zhang, H.; Zheng, X.; Zheng, N.; Shi, W. A Multiscale and Multipath Network with Boundary Enhancement for Building Footprint Extraction fromRemotely Sensed Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8856–8869. [CrossRef]
46. Yang, K.; Zhang, C.; Luo, T.; Hu, L. Automatic Identification Method of Construction and Demolition Waste Based on Deep Learning and GAOFEN-2 Data. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *XLIII-B3-2022*, 1293–1299. [CrossRef]
47. Torres, R.N.; Fraternali, P. AerialWaste Dataset for Landfill Discovery in Aerial and Satellite Images. *Sci. Data* **2023**, *10*, 63. [CrossRef]
48. Zhou, L.; Rao, X.; Li, Y.; Zuo, X.; Liu, Y.; Lin, Y.; Yang, Y. SWDet: Anchor-Based Object Detector for Solid Waste Detection in Aerial Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 306–320. [CrossRef]
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
50. Wang, Z.; Wang, J.; Yang, K.; Wang, L.; Su, F.; Chen, X. Semantic Segmentation of High-Resolution Remote Sensing Images Based on a Class Feature Attention Mechanism Fused with Deeplabv3+. *Comput. Geosci.* **2022**, *158*, 104969. [CrossRef]
51. Tang, Z.; Sun, Y.; Wan, G.; Zhang, K.; Shi, H.; Zhao, Y.; Chen, S.; Zhang, X. Winter Wheat Lodging Area Extraction Using Deep Learning with GaoFen-2 Satellite Imagery. *Remote Sens.* **2022**, *14*, 4887. [CrossRef]
52. Lin, S.; Huang, L.; Liu, X.; Chen, G.; Fu, Z. A Construction Waste Landfill Dataset of Two Districts in Beijing, China from High Resolution Satellite Images. *Sci. Data* **2024**, *11*, 388. [CrossRef]
53. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241, ISBN 978-3-319-24573-7.
54. 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]
55. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239.
56. 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 Computer Vision—ECCV 2018, Munich, German, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 833–851.
57. Lin, S.; Yao, X.; Liu, X.; Wang, S.; Chen, H.-M.; Ding, L.; Zhang, J.; Chen, G.; Mei, Q. MS-AGAN: Road Extraction via Multi-Scale Information Fusion and Asymmetric Generative Adversarial Networks from High-Resolution Remote Sensing Images under Complex Backgrounds. *Remote Sens.* **2023**, *15*, 3367. [CrossRef]
58. Report on the Work of the Government 2023 (Part I). Available online: https://english.beijing.gov.cn/government/reports/202301/t20230129_2908152.html (accessed on 24 November 2023).
59. Huang, L.; Cai, T.; Zhu, Y.; Zhu, Y.; Wang, W.; Sun, K. LSTM-Based Forecasting for Urban Construction Waste Generation. *Sustainability* **2020**, *12*, 8555. [CrossRef]
60. Yuan, L.; Lu, W.; Xue, F. Estimation of Construction Waste Composition Based on Bulk Density: A Big Data-Probability (BD-P) Model. *J. Environ. Manag.* **2021**, *292*, 112822. [CrossRef]

61. Li, Y.; Zhang, X.; Ding, G.; Feng, Z. Developing a Quantitative Construction Waste Estimation Model for Building Construction Projects. *Resour. Conserv. Recycl.* **2016**, *106*, 9–20. [CrossRef]
62. Sivashanmugam, S.; Rodriguez, S.; Pour Rahimian, F.; Elghaish, F.; Dawood, N. Enhancing Information Standards for Automated Construction Waste Quantification and Classification. *Autom. Constr.* **2023**, *152*, 104898. [CrossRef]
63. Han, D.; Kalantari, M.; Rajabifard, A. The Development of an Integrated BIM-Based Visual Demolition Waste Management Planning System for Sustainability-Oriented Decision-Making. *J. Environ. Manag.* **2024**, *351*, 119856. [CrossRef]
64. Wu, Z.; Yu, A.T.W.; Poon, C.S. An Off-Site Snapshot Methodology for Estimating Building Construction Waste Composition—A Case Study of Hong Kong. *Environ. Impact Assess. Rev.* **2019**, *77*, 128–135. [CrossRef]