mathematics

MDPI

Article

WaveSegNet: An Efficient Method for Scrap Steel Segmentation Utilizing Wavelet Transform and Multiscale Focusing

Jiakui Zhong, Yunfeng Xu * and Changda Liu

School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China; 2021114031@stu.hebust.edu.cn.com (J.Z.); 2021114070@stu.hebust.edu.cn.com (C.L.)
* Correspondence: hbkd_xyf@hebust.edu.cn

Abstract: Scrap steel represents a sustainable and recyclable resource, instrumental in diminishing carbon footprints and facilitating the eco-friendly evolution of the steel sector. However, current scrap steel recycling faces a series of challenges, such as high labor intensity and occupational risks for inspectors, complex and diverse sources of scrap steel, varying types of materials, and difficulties in quantifying and standardizing manual visual inspection and rating. Specifically, we propose WaveSegNet, which is based on wavelet transform and a multiscale focusing structure for scrap steel segmentation. Firstly, we utilize wavelet transform to process images and extract features at different frequencies to capture details and structural information in the images. Secondly, we introduce a mechanism of multiscale focusing to further enhance the accuracy of segmentation by extracting and perceiving features at different scales. Through experiments conducted on the public Cityscapes dataset and scrap steel datasets, we have found that WaveSegNet consistently demonstrates superior performance, achieving the highest scores on the mIoU metric. Particularly notable is its performance on the real-world scrap steel dataset, where it outperforms other segmentation algorithms with an average increase of 3.98% in mIoU(SS), reaching 69.8%, and a significant boost of nearly 5.98% in mIoU(MS), achieving 74.8%. These results underscore WaveSegNet's exceptional capabilities in processing scrap steel images. Additionally, on the publicly available Cityscapes dataset, WaveSegNet shows notable performance enhancements compared with the next best model, Segformer. Moreover, with its modest parameters and computational demands (34.1 M and 322 GFLOPs), WaveSegNet proves to be an ideal choice for resource-constrained environments, demonstrating high computational efficiency and broad applicability. These experimental results attest to the immense potential of WaveSegNet in intelligent scrap steel rating and provide a new solution for the scrap steel recycling industry. These experimental results attest to the immense potential of WaveSegNet in intelligent scrap steel rating and provide a new solution for the scrap steel recycling industry.

Keywords: intelligent detection; semantic segmentation; wavelet transform; multiscale focusing; scrap steel

MSC: 68T45

Citation: Zhong, J.; Xu, Y.; Liu, C. WaveSegNet: An Efficient Method for Scrap Steel Segmentation Utilizing Wavelet Transform and Multiscale Focusing. *Mathematics* **2024**, *12*, 1370. https://doi.org/10.3390/math12091370

Academic Editors: Shuo Yu and Feng Xia

Received: 9 April 2024
Revised: 24 April 2024
Accepted: 28 April 2024
Published: 30 April 2024

1. Introduction

The steel industry, a cornerstone of global economic development, plays a pivotal role in driving infrastructure construction and industrial advancement [1]. However, it also represents a sector with high energy and resource consumption, which imposes significant environmental pressure. With increasing awareness of environmental issues and a deeper concern for climate change, the steel industry faces challenges in transformation. Especially in the context of pursuing carbon peaking and carbon neutrality goals, promoting the green and low-carbon transition of the steel industry through technological innovation and process optimization has become a critical issue.

Scrap steel refers to the nonproduct steel waste generated during the steel production process, includes steel from scrapped equipment and components [2]. It is a highly valuable

raw material for the steel manufacturing industry. The complete recyclability of steel ensures that the material input in the production process is fully preserved and can be sustainably recycled [3]. Additionally, due to the magnetic properties of steel, scrap steel can be relatively easily separated and recovered from various waste streams, resulting in a high recovery rate. In steel production, iron ore and recycled scrap steel are primarily used as metal raw materials. Approximately 30% of the iron-containing materials in global steel production come from scrap steel. Through melting and recycling, scrap steel can not only be directly transformed into new steel but can also undergo further refinement in its chemical composition and shape. In blast furnace production processes, scrap steel can be used as a coolant to absorb excess heat generated during the exothermic decarbonization process, and it can also serve as a source of iron material. In the electric arc furnace (EAF) production process, the proportion of scrap steel in the feedstock can reach up to 100%. The recycling of scrap steel plays a crucial role in reducing the carbon emissions of the entire industry and in the reuse of resources. Using every ton of scrap steel in the production process reduces 1.5 tons of carbon dioxide emissions and saves 1.4 tons of iron ore, 0.74 tons of coal, and 0.12 tons of limestone [4].

The recycling and reuse of scrap steel significantly contribute to environmental protection and resource conservation. However, the prevalent reliance on manual operations for scrap steel recycling, characterized by visual inspections and manual measurements, results in low efficiency and a lack of standardized quantitative evaluation criteria. Additionally, the vast diversity and complexity of scrap steel sources, coupled with the variability in shapes, sizes, and materials, further complicates manual assessment. Consequently, the subjective judgments of operators heavily influence scrap steel classification, leading to inconsistencies that can jeopardize the safety of the steel smelting process and compromise the quality of the final products. Furthermore, the recycling work environment, especially in truck loading and unloading areas, is fraught with safety hazards, exposing workers to significant risks. Therefore, the reliance on manual rating of scrap steel no longer meets the development needs of the scrap steel recycling industry.

Current methodologies in scrap steel detection primarily utilize object detection technologies, which effectively identify and localize scrap objects in images. However, the intricate shapes and irregular edges of scrap steel frequently hinder these methods from attaining pixel-level precision in classification, consequently impacting the overall accuracy. To address these issues, we propose the adoption of semantic segmentation technology for the automated identification and classification of scrap steel, thereby enhancing classification accuracy and efficiency while reducing manual labor and associated safety risks. Specifically, we introduce WaveSegNet, a novel framework utilizing wavelet transform and multiscale focusing structure for improved scrap steel segmentation. First, we utilize wavelet transform to process images, extracting features across various frequencies to capture both fine details and broader structural information. Then, we adopt a multiscale focusing mechanism that further refines segmentation accuracy by leveraging feature extraction at different scales. Experimental evaluations on the public Cityscapes dataset and dedicated scrap steel datasets reveal that WaveSegNet surpasses existing advanced models in semantic segmentation efficiency and performance. These findings underscore the immense potential of WaveSegNet in the intelligent classification of scrap steel.

In summary, our main contributions are as follows:

- We propose the application of semantic segmentation within the realm of intelligent classification tasks for scrap steel recycling, aiming to enhance the precision and efficiency of sorting processes through advanced computational methods.
- To better capture the details and structural information of images, we propose the method of downsampling with Daubechies wavelets and upsampling with Haar wavelets to better understand and analyze images.
- We adopt a mechanism of multiscale focusing to further enhance the accuracy of segmentation by extracting and perceiving features at different scales.

- Through extensive experiments, we validate that WaveSegNet exhibits excellent performance in scrap steel segmentation and confirm the effectiveness of various structures.

2. Related Work

2.1. Semantic Segmentation

Semantic segmentation, as one of the significant research directions in computer vision, has found widespread applications in various fields. Its objective is to classify the pixels of input images on a pixel-level basis. Thanks to the powerful capability of CNNs in feature extraction and semantic understanding, FCN was the first to apply CNNs to semantic segmentation [5]. Since then, CNNs-based semantic segmentation techniques have gradually become mainstream. To address the issue of the loss of deep-level pixel positional information and enhance confidence in boundary classification, Ronneberger et al. proposed the symmetrically structured network based on FCN–U-Net [6]. By utilizing skip connections to connect corresponding hierarchical features, U-Net effectively integrates shallow-level details and deep-level semantics. Chen et al. proposed DeepLabV3 [7], which utilizes dilated convolutions to expand the receptive field of the network and achieves multiscale feature extraction by concatenating convolutional kernels with different dilation rates. Subsequently, Chen et al. proposed DeepLabV3+ [8] by introducing depthwise separable convolutions in the ASPP [9] module, leading to further improvement in the performance of semantic segmentation. In recent years, Transformer-based [10–13] approaches have shown great potential in the field of semantic segmentation. These methods utilize self-attention mechanisms to effectively capture long-range dependencies between pixels, surpassing the performance of traditional CNN methods. However, these methods often come with the issue of high computational complexity and significant memory consumption.

2.2. Wavelet Transform in CNNs

In the field of computer vision research, wavelet transform is widely utilized as a powerful tool for multiresolution time–frequency analysis, particularly in image processing such as denoising, enhancement, fusion, and especially compression. Shin Fujieda et al. [14] proposed a novel architecture, Wavelet CNNs, which combines multiscale resolution analysis with CNNs for image classification and annotation tasks. Liu et al. [15] introduced a MultiLevel Wavelet Convolutional Neural Network (MWCNN), which strikes a better balance between receptive field size and computational efficiency, and applied it to image restoration. Wu et al. [16] utilized MWCNN to train a denoiser that effectively removes Cauchy noise and restores blurry images. To address the performance degradation and excessive smoothing issues when dealing with low-resolution images, Huang et al. [17] proposed a wavelet-based CNN method that reconstructs high-resolution images by learning to predict the corresponding high-resolution wavelet coefficients of low-resolution images. Ma et al. [18] presented iWave, an image compression framework based on CNNs that simulates wavelet transform and achieves improved compression performance in both general and specific texture images. However, unlike the aforementioned works, our approach replaces the traditional upsampling and downsampling with wavelet transform.

2.3. Intelligent Scrap Steel Detection

Despite numerous attempts by researchers in the field of intelligent classification for scrap steel, this area is still in its nascent stages of exploration and development. Kim et al. [19] designed an automatic sorting system utilizing image processing techniques, which can automatically sort specified materials from mixtures, particularly extracting copper and other nonferrous metal waste from iron filings mixtures. In [20], researchers proposed an automatic classification system for light metal waste, which measures shape parameters using a 3D imaging camera and performs multivariate analysis with data processing software to achieve accurate classification of waste aluminum and magnesium, potentially replacing traditional manual sorting. Weczorek et al. [21] employed computer vision methods to

extract features such as color and shape from scrap steel images for classification research. Xu et al. [22] utilized machine learning and a nonlinear equal-scale clustering algorithm to achieve an online automated rating of deep-drawn steel product quality. Duan et al. [23] made improvements based on the YOLOv3, realizing the classification and detection of small, light, medium, and heavy scrap steel. Xu et al. [24] proposed CSBFNet for the classification and rating of multiple categories of scrap steel, which exhibits significant advantages in terms of accuracy and fairness compared with manual methods. The intelligent quality inspection system for scrap steel, developed by HBIS Digital Technology Co.Ltd, employs artificial intelligence and machine vision technologies to enable real-time monitoring and automated analysis during the scrap steel unloading process. The system is capable of automatically identifying noncompliant scrap steel, impurities, and foreign objects during unloading and promptly issuing warnings. Additionally, Qingdao Special Iron and Steel Co, Ltd. has developed an automatic scrap steel rating system that constructs a scrap steel rating model capable of analyzing scrap steel images and calculating their similarity to known scrap steel types, thus achieving automated classification [25].

To address the limitations of current methods in achieving pixel-level classification of scrap steel, we propose WaveSegNet. This method is based on wavelet transform and a multiscale focusing structure for scrap steel segmentation. It facilitates classification at the pixel level, offering refined granularity in the categorization of scrap steel and leading to markedly improved results.

3. Scrap Steel Dataset

Currently, an intelligent rating for scrap steel is still in the developmental stage, and there is no publicly available dataset. In this research, we independently collected images of scrap steel and meticulously annotated these images, thereby creating two datasets specifically designed for scrap steel segmentation.

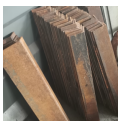
3.1. Simulated Scenario Dataset

In order to obtain accurate and usable data, we used a scrap steel transport car with dimensions of 3 m $\times$ 2 m $\times$ 1 m and purchased various types of scrap steel to simulate the recycling operation site. To ensure the accuracy and completeness of the data, we fixed a Hikvision ball camera with a resolution of 1920 $\times$ 1080 on a bracket at a height of 4.2 m to 4.5 m. This ensures that the camera can overlook the entire car from all angles, capturing a diverse range of scrap steel images. Figure 1 shows some of the collected scrap steel image samples.

Figure 1. The collection site for simulated scenario dataset.

Referencing the national standard for scrap steel classification GB/T 4223-2017 [26], we performed a detailed categorization of the scrap steel images. Table 1 comprehensively lists the categorization criteria along with corresponding examples.

Table 1. Scrap Steel Classification Guidelines.

Category	Description	Example
<3 mm	Thickness of fewer than or equal to 3 mm.	
3–6 mm	Thickness ranging from 3 mm to 6 mm.	
>6 mm	Thickness greater than 6 mm.	
paint	The surface has paint or baked paint coating.	
galvanized	Thickness $\leq$ 2.5 mm and surface coating.	
greasy dirt	The surface is contaminated with oil.	
inclusion	Nonmetallic materials such as rocks, rubber, plastic, sand, etc.	

Based on the aforementioned scrap steel classification standards, we collaborated with workers involved in scrap steel recycling and researchers in related fields to jointly formulate annotation rules applicable to scrap steel images for this study. During the image annotation process using LabelMe, we outlined the boundary region of each piece of scrap steel, determined its grade, and labeled the corresponding label information. These labels include information on the category, shape, and size of the scrap steel. Due to the diverse characteristics of scrap steel, some subjective judgment may be required during the annotation process. After completing the annotation, we conducted quality checks on the annotated data to ensure consistency and accuracy. Figure 2 shows some annotated scrap steel images and their corresponding label mask images.

Figure 2. The original images of the dataset along with their corresponding label mask images.

The simulated scenario dataset consists of 1112 images, each with resolution of 1920 × 1080. We divided the dataset into training and validation sets in an 8:2 ratio, with 890 images used for training and 222 images used for validation. The dataset is stored in the PASCAL VOC [27] format. We conducted a statistical analysis of the number of labels and pixels for each category in the dataset, revealing a total of 81,088 labels, with an average of 73 labels per image. Figure 3a displays the number of labels for each category, while Figure 3b illustrates the pixel count for each category. We observe significant variations in the number of labels across different categories, as well as distinct differences in the proportion of pixels occupied by different category labels. Among them, the category with the highest number of labels is ">6 mm" which accounts for 77.34% of the total labels. In terms of pixel count, ">6 mm" also has the highest proportion, representing 60.80% of the total pixels. In contrast, the category with the fewest samples is "<3 mm", which accounts for only 1.27% of the total labels. The category with the smallest pixels is "inclusion", representing a mere 0.53% of the total annotated pixels.

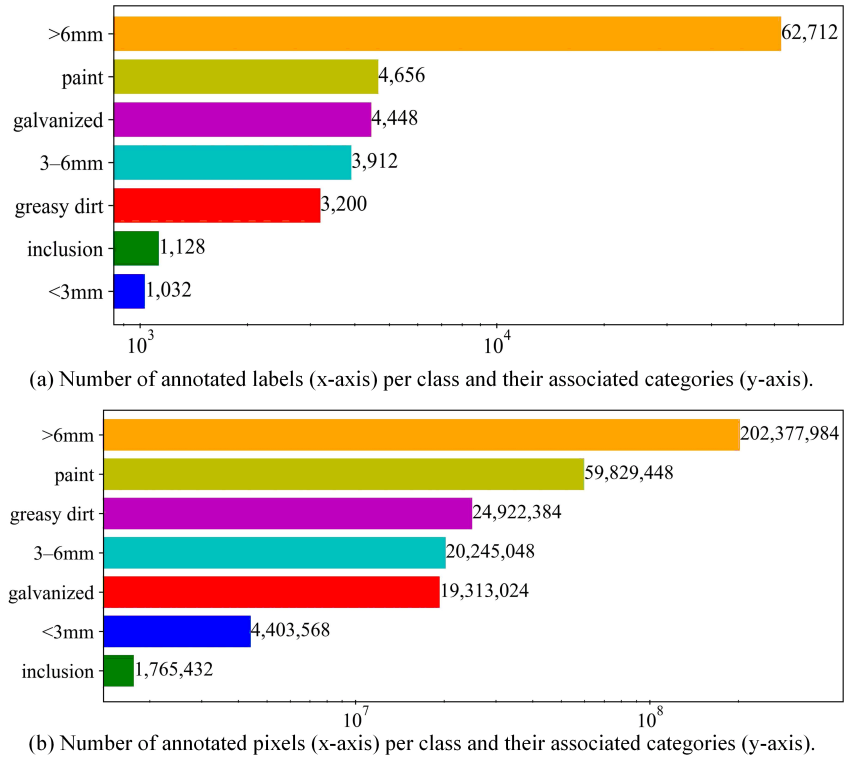
In the real-world environment of scrap steel recycling, capturing images of scrap steel is often subject to a variety of external interferences, such as changes in lighting, physical obstructions, and various weather conditions. These factors not only increase the complexity of image processing but also make the identification of scrap steel features more challenging. To enhance the model's adaptability to these real-world disturbances, we employed data augmentation techniques to increase the diversity of the scrap steel image dataset. The specific data augmentation techniques applied are as follows:

Rotation: By randomly rotating scrap steel images, we simulate the visual effect of scrap steel at different viewing angles, training the model to accurately identify scrap steel from all directions.

Occlusion: Introducing random occlusion elements into scrap steel images simulates potential line-of-sight obstructions that may occur on-site, enhancing the model's ability to recognize features when partial information is obscured.

Shadowing: Applying shadows of varying intensities and directions to simulate the appearance of scrap steel under changing lighting conditions, improving the model's adaptability to light variations.

Noise: Incorporating various types of random noise (e.g., Gaussian noise, salt-and-pepper noise) into the scrap steel images to simulate potential visual disturbances encountered in real environments.

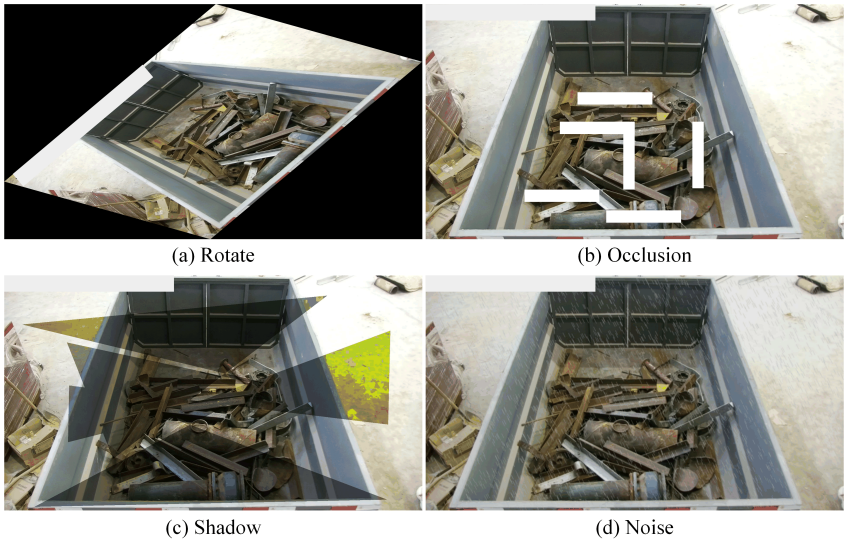


(a) Number of annotated labels (x-axis) per class and their associated categories (y-axis).

(b) Number of annotated pixels (x-axis) per class and their associated categories (y-axis).

Figure 3. Detailed information about simulated scenario dataset. The same category is the same color.

As shown in Figure 4, the adoption of the aforementioned data augmentation strategies has generated a more diverse dataset, thus significantly improving the model's robustness when dealing with complex disturbances in real-world scenarios.



(a) Rotate (b) Occlusion

(c) Shadow (d) Noise

Figure 4. Images of scrap steel following various data augmentation.

3.2. Real-World Scenario Dataset

While simulated data can provide preliminary insights, they may not fully capture the complexity of real scrap steel recycling sites, potentially leading to a lack of generalization capability in practical applications. To bridge this gap, our research team established a collaboration with a steel plant to collect the required scrap steel images directly. To ensure the high quality of the data, we employed three Hikvision 8-megapixel network

cameras for image collection, which can capture high-definition images with a resolution of 3840 × 2160.

To optimize viewing angles and expand coverage, these cameras were installed on supports ranging from 10.2 to 10.5 m in height at the scrap steel recycling site. This setup not only provides an overhead view of the entire site but also captures images from multiple angles, significantly enhancing the diversity and comprehensiveness of the data. The methodology for image collection and the site configuration are detailed in Figure 5.

Figure 5. The collection site for the real-world scenario dataset.

In the practical scenarios of scrap steel recycling operations, we encounter a wide variety of scrap steel types, each with its unique characteristics and properties. To achieve more accurate identification and description of these categories, we have adopted a more detailed classification method, subdividing scrap steel into 19 specific categories. This refined classification not only allows us to gain a deeper understanding of the diversity and complexity of scrap steel but also provides more precise guidance for the effective recycling and processing of scrap steel. In Table 2, we present several common types of scrap steel found in real recycling settings, along with their classification standards.

Table 2. Scrap Steel Classification Guidelines.

Category	Description	Example
overweight	Single piece weight > 700 KG.	
airtight	Container isolated from external environment.	
scattered	Fine shredded steel powder, rust, iron filings, etc.	
cast_iron	Cast iron with a carbon content ranging from 2% to 6.67%.	
ungraded	There is significant contamination, corrosion on the surface, or defects in size and shape.	

The real-world dataset, akin to the simulated scenario dataset, consists of 1340 carefully annotated scrap steel images, each with a resolution of 3840 × 2160 pixels. These images

are split into training and validation sets following an 8:2 ratio, allocating 1088 images for training the model and 272 for validating its performance. The data are meticulously organized and stored in the PASCAL VOC format, with the dataset featuring a total of 134,584 labels, averaging about 100 labels per image.

As depicted in Figure 6, there is a notable disparity in the number of labels across different categories. Remarkably, the "3–6 mm" accounts for the largest proportion, making up 43.28% of the overall label count, while the "steel_bar(>2 m)", with the fewest labels, accounts for only 0.01% of the total. Additionally, Figure 7 depicts the variation in pixel counts for annotations across various categories, revealing significant differences in the volume of pixel annotations among them. Interestingly, while the "3–6 mm" had the highest number of labels, it did not lead in terms of annotated pixel volume. This trend was also evident in other categories, with the ">6 mm" category comprising 43.77% of the total annotated pixel volume.

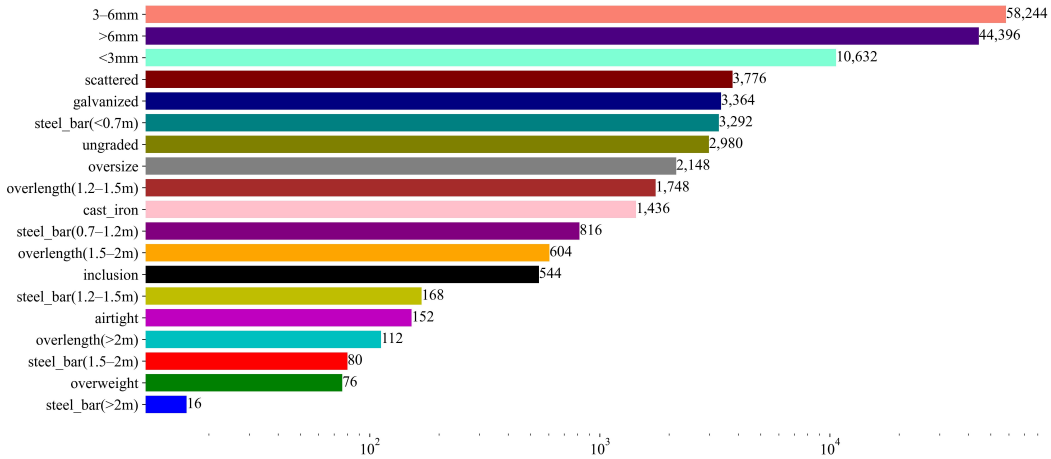


Figure 6. The number of labels corresponding to each category of scrap steel.

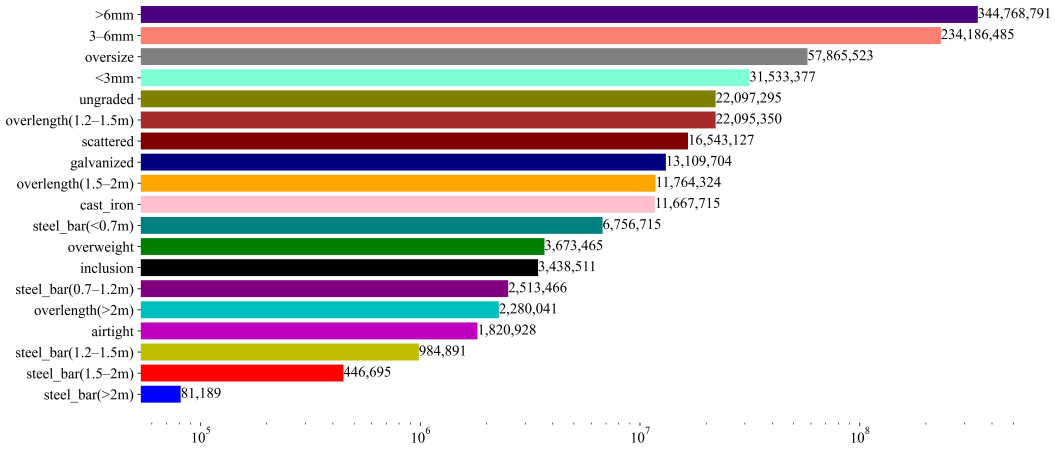


Figure 7. The number of pixels corresponding to each category of scrap steel.

In the images collected, not only is the wagon's interior filled with scrap steel, but there are also extraneous piles of scrap steel around the exterior, causing disruption to the identification process. The majority of the scrap steel and waste materials within the wagon consist of small, densely arranged pieces that are tightly stacked, resulting in considerable overlap among the targets. This overlap leads to significant mutual obstruction, preventing a complete exposure of all scrap steel targets. Furthermore, even among scrap steel of the same category, there is a notable lack of uniformity in shape, which further complicates the identification process.

4. Methods

In this section, we provide a comprehensive overview of the WaveSegNet that we have proposed for scrap steel image segmentation. As illustrated in Figure 8, WaveSegNet employs an encoder-decoder architecture similar to previous works. In order to design a powerful backbone network capable of handling complex scrap steel scene segmentation, we incorporate a four-stage feature hierarchy to generate feature maps at different scales. Each stage consists of a downsampling module and multiple MultiScale Focusing Convolution Attention Blocks (MSFCABs).

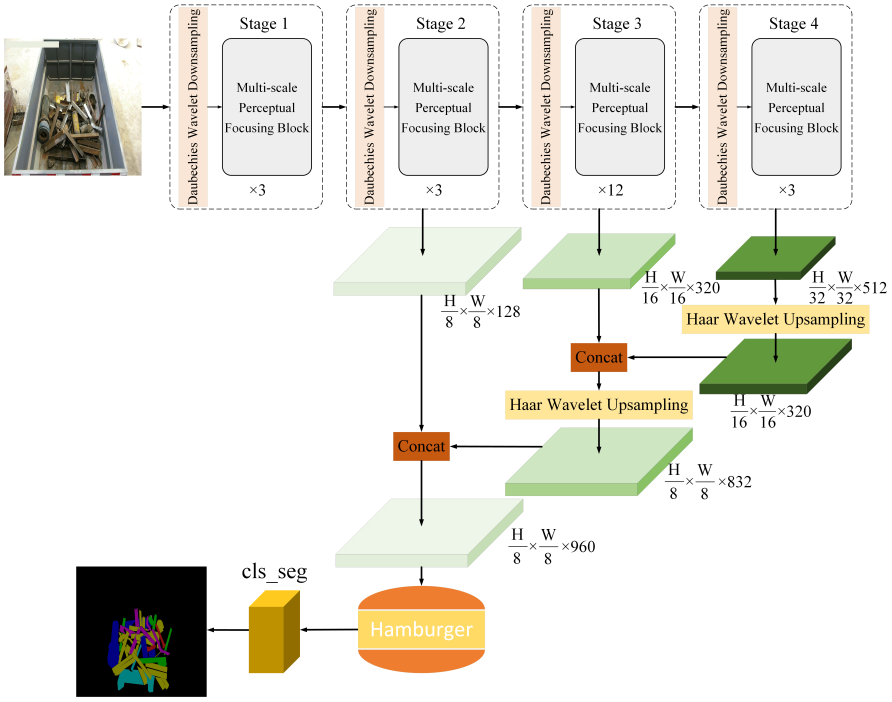


Figure 8. The architecture of WaveSegNet.

4.1. Encoder

4.1.1. MultiScale-Focusing-Based Self-Attention

Research indicates that short-term context and long-term context both play important roles in visual modeling [28,29]. Short-term context provides information about local details and spatial relationships, helping to identify low-level features such as edges and textures and capture relationships between local objects. Long-term context, on the other hand, provides global semantic and contextual information, which helps to understand the relationship between targets and the overall. To comprehensively consider the roles of short-term context and long-term context, we propose MSPF, as shown in Figure 9.

Feature Map Projection. For the input feature map $F \in \mathbb{R}^{C\times H\times W}$, we first use DWConv for mapping to generate the feature map $F_{kv} \in \mathbb{R}^{2\times C\times H\times W}$. It aims to extract a richer and more diverse feature representation from the original feature map. Next, we partition F_{kv} into $F_k, F_v \in \mathbb{R}^{C\times H\times W}$, along the channel dimension. The above operation can be represented as

$$F_k, F_v = Split(\text{DW-Conv}_{5\times 5}, [C, C], dim = 1) \quad (1)$$

where *Split* represents the function *torch.split()* , DW-$\text{Conv}_{5\times5}$ represents a depth-wise separable convolution with a kernel size of 5 × 5, stride of 1, and the number of groups is C. $[C, C]$ indicates the size or splitting positions of each subtensor after splitting, *dim* represents the split dimension. F_k and F_v are the output feature maps.

MultiScale Perception Aggregation. To obtain visual tokens of different grains, we perform multiscale aggregation on the feature maps F_k and F_v. We use patches of different sizes to generate features of the same resolution in order to capture feature information

at different scales. Specifically, we perform multiscale convolution on the feature maps by using convolutional kernels with different receptive fields. For the feature map F_k, we employ depth-wise separable convolutions with kernel sizes of 7×7, 11×11, and 21×21. This enables us to capture feature maps $F_{ki} \in \mathbb{R}^{C \times H \times W}$ with different scale feature information, where $i \in \{1, 2, 3\}$. This operation can be represented as

$$F_{ki} = focus(F_k, Scale_i \times Scale_i) \tag{2}$$

where $focus$ represents a depth-wise separable convolution with a stride of 1 and the number of groups equals C. $Scale_i \times Scale_i$ represents the kernel size, and F_{ki} represents the generated feature map.

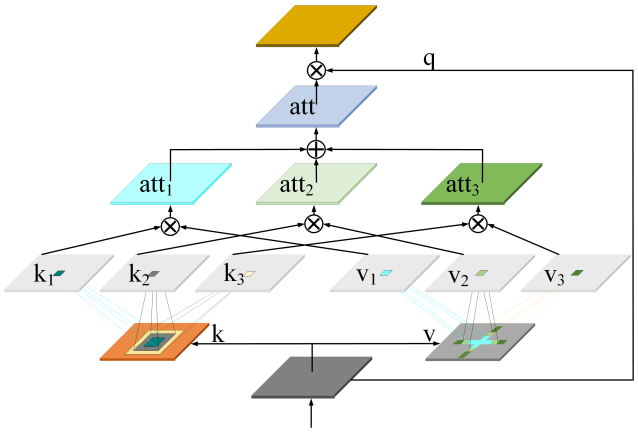


Figure 9. The architecture of MSPF.

Compared with traditional convolution, stripe convolution is more lightweight and better suited for extracting features from objects with stripe-like shapes. Scrap steel often has a long and striped shape, and using stripe convolution can better capture the features of scrap steel, thereby improving the accuracy of model recognition. For F_v, we utilize three sets of stripe convolutions instead of standard convolution to obtain feature maps $F_{vi} \in \mathbb{R}^{C \times H \times W}$. For instance, we replace the convolutional kernel size of $K \times K$ with a set of convolutional kernels of size $K \times 1$ and $1 \times K$. This can be expressed as

$$F_{vi} = focus(focus(F_v, Scale_i \times 1), 1 \times Scale_i) \tag{3}$$

Through multiscale aggregation, we are able to obtain fine-grained and coarse-grained visual representations, thereby capturing feature information at different scales more effectively.

Local Feature Interaction–Global Feature Fusion. In this step, our goal is to extract information about structure and local relationships from features at different levels and combine them with a global context transformer. Specifically, we first take the element-wise multiplication and addition of the local $F_k, F_v \in \mathbb{R}^{C \times H \times W}$ at different granularity levels, obtaining the feature map $F_{att} \in \mathbb{R}^{C \times H \times W}$. This not only achieves feature aggregation across different levels but also preserves the detailed information and encodes the relationships between local features. Then, the global context transformer is used to globally attend to and allocate weights to the feature map F_{aat}. Mathematically, this process can be summarized as

$$F_{out} = F \otimes G(\sum_{i=0}^{3}(F_{ki} \otimes F_{vi})) \in \mathbb{R}^{C \times H \times W} \tag{4}$$

The function G represents a convolution with a kernel size of 1×1 and a stride of 1. The symbol $\otimes$ denotes element-wise multiplication. F_{out} represents the output feature map.

Through this, we can fully leverage the advantages of features at different levels, enhancing the ability to understand overall semantics and local relationships.

4.1.2. Daubechies Wavelet Downsampling

In segmentation tasks, traditional CNNs often employ pooling operations, such as max pooling or average pooling, or use strided convolutions for downsampling to reduce the size of feature maps and extract more salient features. However, these methods can lead to the loss of crucial information, particularly affecting fine-grained details such as boundaries, textures, and small-sized object details. Wavelet transforms enable multiscale, lossless signal decomposition, effectively retaining crucial image details often lost in conventional downsampling techniques. Our method diverges by significantly increasing the number of channels within the feature maps and concurrently reducing their spatial resolution through the application of wavelet transform. This approach not only maintains the integrity of the image information but also enhances the ability to discern more nuanced and discriminative features, thereby elevating the overall performance of segmentation tasks.

Drawing inspiration from the lossless information transformation method [30], we incorporate the Daubechies wavelet transform in the downsampling module, as illustrated in Figure 10. We refer to this as Daubechies Wavelet Downsampling (DWD), which consists of two components: lossless feature encoding and feature representation learning. For an input feature map of size H × W × C, after the Daubechies wavelet transform, we obtain four components: the approximation (low-frequency) component (A), as well as the horizontal component (H), vertical component (V), and diagonal detail (high-frequency) component (D). The size of each component is $H/2 \times W/2 \times C_{in}$. The Daubechies wavelet transform can encode partial information from the spatial dimension to the channel dimension without any loss of information. Through this transformation, we can concatenate four components along the channel direction, resulting in a feature map of size $H/2 \times W/2 \times 4C_{in}$. Subsequently, through the feature representation learning part, we aim to filter out redundant information as much as possible, enabling subsequent layers to learn more effective representative features. Furthermore, the channel number of the feature map can be adjusted by manipulating the representation learning block.

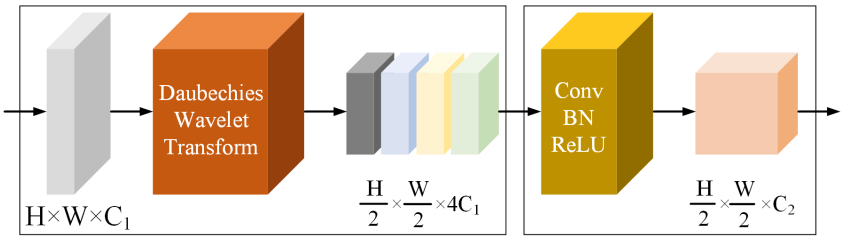


Figure 10. The architecture of Daubechies Wavelet Downsampling.

4.2. Decoder

The primary function of the decoder is to transform abstract high-level semantic features into more specific and interpretable forms. During this transformation process, the decoder utilizes feature concatenation and fusion from different hierarchies to achieve pixel-level classification and segmentation.

As depicted in Figure 11, the decoder employed in this study achieves its functionality by aggregating features from the last three stages. It utilizes wavelet transform upsampling to sample and aggregate the features from different stages. For an input feature map of size H × W × C, a learned feature representation generates a feature map of size H × W × 4C. This feature map is then split along the channel direction into four feature components of size H × W × C. Subsequently, Haar wavelet inverse transform is applied to recombine these components into a feature map of size 2H × 2W × C.

For the feature maps sampled and aggregated at different stages, we use a lightweight Hamburger [31] to further model the global context, resulting in a more comprehensive and representative feature representation. As depicted in Figure 12, the Hamburger consists of three main parts: the middle layer, or the ham layer, which is dedicated to matrix factorization, and the upper and lower layers, collectively known as the bread layers,

which perform linear transformations. The lower bread layer transforms the input features $Z \in \mathbb{R}^{d_z \times n}$ into a new feature space $\mathbb{R}^{d \times n}$, facilitated by the weight matrix $W_l \in \mathbb{R}^{d \times d_z}$. The ham layer then utilizes matrix factorization, denoted by **M**, to further process these features, extracting a low-rank signal subspace and uncovering latent structures within the data. The upper bread layer maps the features processed by the ham layer back to the original feature space through another linear transformation, represented by the weight matrix $W_u \in \mathbb{R}^{d_z \times d}$. By integrating linear transformations with matrix factorization, the Hamburger effectively extracts features and learns representations.

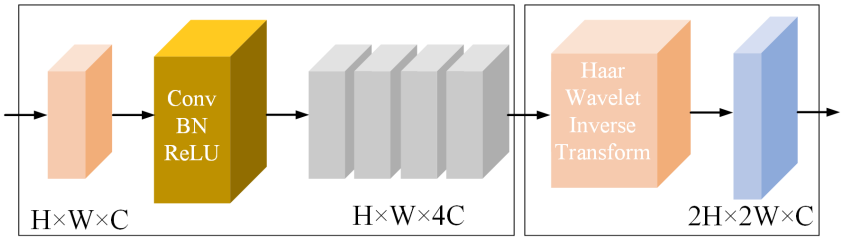


Figure 11. The architecture of Haar wavelet upsampling.

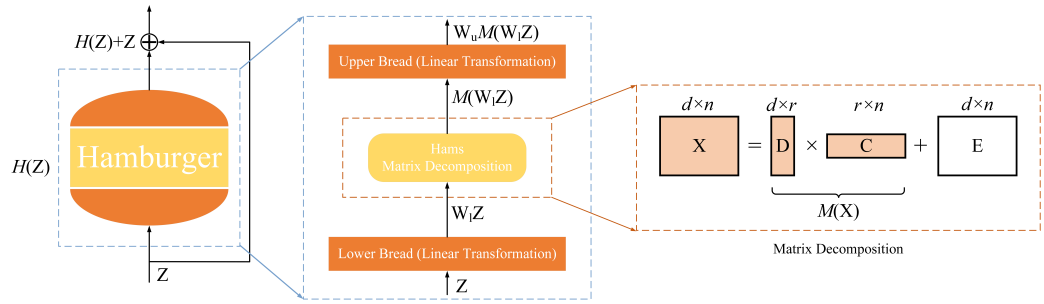


Figure 12. The architecture of the Hamburger.

Unlike most decoders, we only aggregate feature information from the last three stages. This is because the first stage contains a significant amount of low-level information that can negatively impact performance. Additionally, it introduces a heavy computational burden. In the experimental section, we demonstrate that WaveSegNet outperforms other advanced methods.

5. Experiments

To evaluate the performance of WaveSegNet, we carried out comprehensive experiments across three distinct datasets. These experiments were performed on Ubuntu 20.04 using the PyTorch. For training and validation, we utilized two NVIDIA GeForce RTX 4090. Detailed specifications of the experimental setup are provided in Table 3.

Table 3. Experimental environment information.

Experimental Configuration	Detailed Information
Operating System	Ubuntu 20.04
Motherboard	ROG MAXIMUS Z790 HERO
CPU	13th Gen Intel(R) Core(TM) i9-13,900 K
GPU	NVIDIA GeForce RTX 4090 (24 G) × 2
RAM	64 GB
Storage space	6 TB
GPU Driver Version	520.56.06
CUDA Version	11.8
Python Version	3.8.13
PyTorch Version	1.13.0

To ensure fair comparison, all experiments were implemented using the MMSegmentation library, with backbone networks of different models pretrained on ImageNet.

The AdamW [32] optimizer was employed with an initial learning rate of 6×10^{-5} and a weight decay rate of 0.01. The experiments leveraged a linear learning rate decay scheduler alongside a linear warm-up strategy over 1500 iterations. A suite of data augmentation techniques were employed, encompassing random resizing (within a scale range of 0.5–2.0), random horizontal flips, and random cropping. For the simulated scenario dataset, the input size was cropped to 512 × 512, whereas the real-world scenario dataset was cropped to 1024 × 1024. Further details on the experimental parameter settings are provided in Table 4. Both single-scale (SS) and multiscale (MS) flip testing methodologies were employed to ensure a balanced evaluation.

Table 4. Experimental parameter setting.

Parameter	Simulated Scenario	Real-World	Parameter Description
img_scale	2048 × 512	2048 × 1024	Image resizing dimensions
ratio_range	(0.5, 2.0)	(0.5, 2.0)	Range for image scaling ratios
crop_size	512 × 512	1024 × 1024	Image cropping size
cat_max_ratio	0.75	0.75	Maximum ratio for object cropping
prob	0.5	0.5	Image flip probability
batch size	16	8	Number of samples per batch
max_iters	40 k	160 k	Training iterations
optimizer	AdamW	AdamW	Type of optimizer
betas	(0.9, 0.999)	(0.9, 0.999)	Momentum parameters for AdamW optimizer
lr	6×10^{-5}	6×10^{-5}	Learning rate
warmup	linear	linear	Learning rate warm-up method
warmup_iters	1500	1500	Iterations for learning rate warm-up
warmup_ratio	1×10^{-6}	1×10^{-6}	Minimum learning rate ratio during warm-up
min_lr	0.0	0.0	Minimum learning rate
weight_decay	0.01	0.01	Weight-decay coefficient

5.1. Performance Evaluation Metrics

To thoroughly assess model performance, we employ standard metrics such as Parameters (Params), Floating Point Operations (FLOPs), and mean Intersection over Union (mIoU). Specific definitions and methods for calculating these are provided below.

Params refers to the number of learnable parameters within the model, serving as a crucial indicator of model complexity. Assuming a neural network consists of L layers, with the dimension of the weight parameters at layer l denoted by $W_l \in \mathbb{R}^{C_{\text{out}}^l \times C_{\text{in}}^l \times H^l \times W^l}$, and the dimension of the bias parameters by $b_l \in \mathbb{R}^{C_{\text{out}}^l}$, then the total number of parameters for this network can be expressed as follows:

$$Params = \sum_{l=1}^{L} (C_{\text{out}}^l \times C_{\text{in}}^l \times H^l \times W^l + C_{\text{out}}^l) \tag{5}$$

where C_{in}^l and C_{out}^l, respectively, represent the number of input channels and output channels at layer l, and H^l and W^l denote the height and width of the channels at layer l.

FLOPs measure the number of floating-point operations required to execute the network model once. This metric is commonly used to evaluate models' computational efficiency and processing speed. Assuming a neural network has L layers, with the input feature map size of layer l denoted as $H_{\text{in}}^l \times W_{\text{in}}^l$, the output feature map size as $H_{\text{out}}^l \times W_{\text{out}}^l$, the kernel size of the convolutional layer as $H_{\text{k}}^l \times W_{\text{k}}^l$, the number of input channels as C_{in}^l, and the number of output channels as C_{out}^l; the total FLOPs can be represented as follows:

$$FLOPs = \sum_{l=1}^{L} 2 \times C_{\text{in}}^l \times C_{\text{out}}^l \times H_{\text{k}}^l \times W_{\text{k}}^l \times H_{\text{out}}^l \times W_{\text{out}}^l \tag{6}$$

The mIoU is a crucial metric for evaluating semantic segmentation. It determines the average ratio of overlap between predicted results and ground truth across all categories. Further, mIoU is divided into single-scale and multiscale evaluations. The mIoU(SS) can be represented as follows:

$$\text{mIoU(SS)} = \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{7}$$

where n_c is the number of categories, TP_i is the number of true positives for category i, FP_i is the number of false positives for category i, and FN_i is the number of false negatives for category i.

The mIoU(MS) can be represented as follows:

$$\text{mIoU(MS)} = \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{\sum_j TP_{i,j}}{\sum_j (TP_{i,j} + FP_{i,j} + FN_{i,j})} \tag{8}$$

where $TP_{i,j}$ is the number of true positives for category i at scale j, $FP_{i,j}$ is the number of false positives for category i at scale j, and $FN_{i,j}$ is the number of false negatives for category i at scale j.

5.2. Semantic Segmentation on Scrap Steel

5.2.1. Simulated Scenario Dataset

As shown in Table 5, we evaluated the key metrics of the different models and analyzed their mIoU on the single scale (SS) and multiscale (MS). Among these models, WaveSegNet exhibits superior segmentation performance across different scales. Firstly, WaveSegNet attained mIoUs of 73.1% and 73.7% on the single scale (SS) and multiscale (MS), respectively, outperforming other models. The mIoU is pivotal in assessing the precision of a model's segmentation outcomes, highlighting WaveSegNet's superior ability to accurately interpret the semantic content of images and achieve better segmentation outcomes. Secondly, although the performance improvement of WaveSegNet is not markedly significant compared with the Swin and ConvNeXt, its parameter count is only half of those models, and its computational demand is reduced to one-sixth. This indicates a higher efficiency in model architecture and parameter utilization for WaveSegNet, which is crucial for real-time segmentation tasks with limited computational resources. Even compared with Segformer, which has a similar scale of parameters and computational requirements, WaveSegNet demonstrates superior performance.

Table 5. Semantic segmentation result on simulated scenario dataset. The arrows indicate the desirable direction for each metric: (↓) for Params and FLOPs, indicating lower is better; (↑) for mIoU, indicating higher is better.

Model	Params (M) ↓	FLOPs (G) ↓	mIoU (SS) ↑	mIoU (MS) ↑
WaveSegNet	34.1	**321**	**73.1**	**73.7**
DeepLabv3+ [8]	43.6	1403	71.9	72.4
MPViT [33]	105.2	2365	69.6	71.4
Segformer [34]	**24.7**	325	71.9	72.5
Swin [35]	59.8	1879	72.2	72.4
ConvNeXt [36]	60.1	1868	72.5	73.4

Bold values indicate optimal quantities.

In Figure 13, the segmentation results of different models on scrap steel images are presented, with WaveSegNet demonstrating exceptional performance in delineating object details. It precisely identifies and segments various components within the scrap steel images, capturing subtle features and edges.

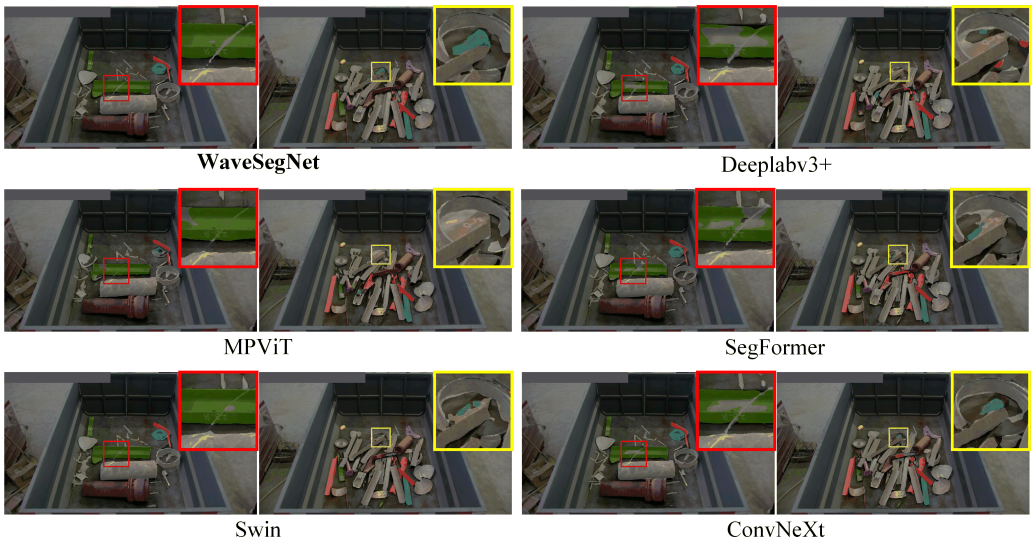


Figure 13. The comparison of performance of different models on simulated scenario dataset.

5.2.2. Real-World Scenario Dataset

As demonstrated in Table 6, WaveSegNet achieved the highest IoU in both single-scale (SS) and multiscale (MS) evaluations, with results of 69.8% and 74.8%, respectively. This underscores WaveSegNet's exceptional capability in performing scrap steel segmentation tasks in real-world scenarios. Compared with DeepLabv3+ and Segformer, WaveSegNet demonstrated substantial enhancements in segmentation efficiency, despite possessing a parameter count that is comparable. Specifically, within the single-scale evaluation, WaveSegNet exceeded the performance of DeepLabv3+ and Segformer by margins of 4.6% and 4.1%, respectively. In the multiscale assessment, the respective advancements stood at 8.5% and 4.0%. It is particularly noteworthy that WaveSegNet not only showcases pronounced advantages in terms of parameter count and computational efficiency relative to Swin and ConvNeXt but also excels in segmentation performance. Additionally, the experimental results suggest that MPViT may not be suitable for scrap steel segmentation and rating. In summary, WaveSegNet not only showcases exemplary segmentation performance but also stands out in terms of parameter efficiency and computational economy, revealing its substantial potential in scrap steel image segmentation.

Table 6. Semantic segmentation results on the real-world dataset. The arrows indicate the desirable direction for each metric: (↓) for Params and FLOPs, indicating lower is better; (↑) for mIoU, indicating higher is better.

Model	Params (M) ↓	FLOPs (G) ↓	mIoU (SS) ↑	mIoU (MS) ↑
WaveSegNet	34.1	**322**	**69.8**	**74.8**
DeepLabv3+	43.6	1404	65.2	66.3
MPViT	105.2	2368	57.3	58.5
Segformer	**27.4**	420	65.7	70.8
Swin	59.8	1880	64.1	66.3
ConvNeXt	60.1	1869	65.6	69.0

Bold values indicate optimal quantities.

Figure 14 showcases the IoU and Accuracy (Acc) of WaveSegNet for different categories of scrap steel in the segmentation task. Notably, the model demonstrates exceptional performance in eliminating background interference, with the IoU for the background reaching 98.86% and an Acc of 99.56%. For the majority of scrap steel categories, both IoU and Acc exceed 80%, with some categories even surpassing 90%, which emphatically validates the effectiveness and reliability of the model proposed in this study. However, it is important to highlight that certain scrap metal categories, such as steel_bar (2 m)

and steel_bar (0.7–1.2 m), exhibit comparatively weaker performance. This is primarily attributed to two reasons: first, the relatively limited sample size of these specific categories in the dataset poses challenges for the model; second, these categories of scrap steel often present as elongated shapes, which lack distinctive features, thereby complicating the segmentation task. In contrast, despite having only 76 annotated samples for the overweight, the model achieves optimal segmentation performance due to its pronounced distinguishing features.

Figure 14. Segmentation results of different categories of scrap steel.

In Figure 15, we present the segmentation results of various models on the real-world scenario dataset. Notably, WaveSegNet stands out in its handling of object details.

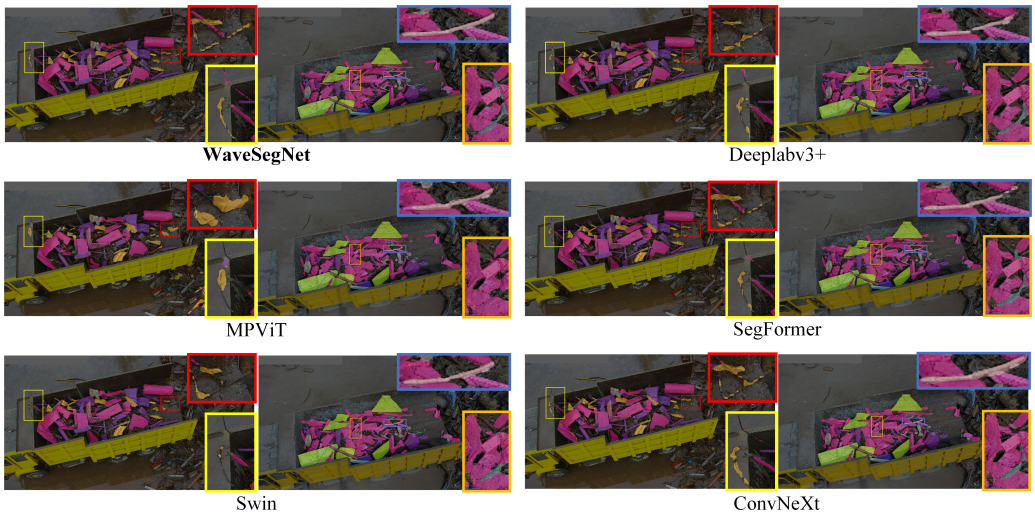


Figure 15. The comparison of performance of different models on the real-world dataset.

In brief, the excellent performance of WaveSegNet in scrap steel segmentation can be attributed to the following aspects:

- Lossless wavelet transform: WaveSegNet uses lossless wavelet transform for upsampling and downsampling, ensuring the integrity of image information and accurately segmenting the boundaries of scrap steel.
- Multiscale perception focusing: WaveSegNet adopts a multiscale perception focusing mechanism to concentrate the attention on the scrap steel area, thereby reducing the impact of background interference.
- Task-specific customization: WaveSegNet is tailor-made and fine-tuned for the specific task, aligning with the unique features and demands of scrap steel images.

5.3. Semantic Segmentation on Cityscapes

To facilitate a more comprehensive benchmark comparison with other models, we additionally selected the widely used Cityscapes dataset, which is a public dataset for semantic segmentation tasks for benchmark testing. The Cityscapes [37] dataset comprises high-resolution images with annotations for 19 different classes. Specifically, it consists of 5000 finely annotated images, out of which 2975 were allocated for model training, 500 for validation, and 1525 for testing. The detailed parameter settings for the experiment are consistent with those used in the real-world dataset in the experiment shown in Table 4.

As shown in Table 7, we evaluated several semantic segmentation models and compared them with WaveSegNet in terms of key indicators, such as model parameters, computational complexity (FLOPs), and mIoU. We found that WaveSegNet has a slightly higher parameter count than Segformer, but has lower computational complexity and achieves a higher mIoU of 81.8%. Compared with Deeplabv3, WaveSegNet achieves a performance improvement of 2.5% with a reduction in parameter count by half and a reduction in computation by 85%. Compared with Transformer-based Swin and CNN-based ConvNeXt, WaveSegNet achieves superior performance using less than 60% of the parameter count and 20% of the computation. In summary, WaveSegNet exhibits significant advantages in terms of parameter size, computational complexity, and mIoU, and achieves the best performance in the Cityscapes semantic segmentation task.

Table 7. Semantic segmentation result on Cityscapes validation set. The arrows indicate the desirable direction for each metric: (↓) for Params and FLOPs, indicating lower is better; (↑) for mIoU, indicating higher is better.

Model	Params (M) ↓	FLOPs (G) ↓	mIoU ↑
WaveSegNet	34.1	**322**	**81.8**
Deeplabv3 [7]	68.1	2157	79.3
Deeplabv3+	43.6	1414	80.1
Segformer	**27.5**	420	81.0
Swin	59.8	1871	79.5
ConvNeXt	61.1	1869	80.7

Bold values indicate optimal quantities.

5.4. Ablation Study

In this section, we conduct ablation studies on key components in WaveSegNet to verify the effectiveness of these designs.

Focusing Branch. Multiscale perceptual aggregation includes three branches. According to the experimental results in Table 8, when we remove the smallest 7 × 7 focusing branch, the performance slightly decreases, with a decrease of 0.2% on Cityscapes and a decrease of 0.2% on the scrap steel dataset. However, these performance losses are not significant. On the other hand, if we remove the largest 21 × 21 focusing branch, the performance on Cityscapes would decline from 81.8% to 80.5%, while performance on the scrap steel dataset would decrease from 73.1% to 71.9%. This shows that the largest scale branch contributes more significantly to performance, and removing it leads to greater performance loss. By removing different scales of focusing branches, we can find that each scale of focusing branch contributes to the final performance.

Table 8. The ablation experiment results of WaveSegNet on Cityscapes and the simulated scenario dataset. Symbols used: (✓) indicates inclusion of a feature, (×) indicates exclusion.

Ablation	Variant		Cityscapes			Simulated Scenario Dataset		
			Params (M)	FLOPs (G)	mIoU	Params (M)	FLOPs (G)	mIoU
Baseline	WaveSegNet		34.14	321.52	81.8	34.13	321.36	73.1
Focus Branch	Remove 7 × 7 branch		33.75	316.80	81.6	33.74	313.92	72.9
	Remove 11 × 11 branch		33.27	310.96	81.2	33.26	310.80	72.5
	Remove 21 × 21 branch		31.25	286.40	80.5	31.24	286.16	71.9
	Downsampling	Upsampling						
Wavelet Transform	✓	×	30.30	296.56	81.6	30.30	296.40	73.0
	×	✓	35.21	327.76	81.5	35.21	327.52	72.7
	×	×	31.38	302.80	81.2	31.37	302.56	72.5

Wavelet Transform. To validate the impact of the proposed wavelet transform downsampling and upsampling on performance, we conducted ablation studies by replacing them with traditional stride convolution downsampling and linear interpolation upsampling. The experimental results are summarized in Table 8. We attempted to replace the Haar wavelet transform upsampling in the decoder with a traditional bilinear interpolation upsampling. However, it can be observed that on different datasets, the segmentation performance decreased to varying degrees. For the encoder, we replaced the Daubechies wavelet downsampling with traditional stride convolution. Similarly, we observed a decrease in segmentation performance of 0.3% on Cityscapes and a decrease of 0.4% on the scrap steel dataset. When both the encoder and decoder used traditional methods instead of wavelet transforms, the results decreased more significantly. This suggests that wavelet transforms can better preserve details and semantic information during upsampling and downsampling.

6. Conclusions and Future Works

6.1. Conclusions

In this study, we proposed the WaveSegNet based on wavelet transform and multiscale focusing structure for scrap steel segmentation. By applying wavelet transform to the images, we are able to extract features at different frequencies, effectively capturing the fine details and structural information of the images. Furthermore, the introduction of a multiscale focusing mechanism enhances the accuracy by enabling the extraction and perception of features across different scales. Through experiments conducted on publicly dataset Cityscapes and our custom-built scrap steel dataset, we demonstrate that WaveSegNet exhibits superior performance and efficiency in the domain of semantic segmentation, surpassing other advanced models. These experimental results confirm the substantial potential of WaveSegNet for intelligent scrap steel rating and offer a new solution for the scrap steel recycling industry.

6.2. Limitations and Future Work

We acknowledge that our dataset may not fully capture the diversity and complexity present in scrap steel. Consequently, we warmly invite further exploration by fellow researchers in this field. Our team is committed to sharing our findings and insights to foster a more complete understanding of intelligent scrap steel recycling and to collectively advance the state of research in this vital area. In the future, we plan to continue expanding the scrap steel dataset, collecting a more comprehensive set of images that covers a broader range of types and scenarios. Additionally, we aim to explore further optimization strategies to enhance the performance of WaveSegNet in scrap steel image processing.

Furthermore, we plan to collaborate with additional enterprises to apply this method in practical production.

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

Funding: This research was funded by Supported projects of key R & D programs in Hebei Province (No.21373802D) and Artificial Intelligence Collaborative Education Project of the Ministry of Education (201801003011).

Data Availability Statement: To protect the privacy of our data sources and stakeholders' interests amidst industry competition and commercial factors, we cannot make the dataset publicly available. The data presented in this study are available from the corresponding author upon reasonable request.

Acknowledgments: The GPU server in this article is jointly funded by Shijiazhuang Wusou Network Technology Co., Ltd. and Hebei Rouzun Technology Co., Ltd.

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

References

1. Akram, R.; Ibrahim, R.L.; Wang, Z.; Adebayo, T.S.; Irfan, M. Neutralizing the surging emissions amidst natural resource dependence, eco-innovation, and green energy in G7 countries: Insights for global environmental sustainability. *J. Environ. Manag.* **2023**, *344*, 118560. [CrossRef]
2. Ma, Y.; Wang, J. Time-varying spillovers and dependencies between iron ore, scrap steel, carbon emission, seaborne transportation, and China's steel stock prices. *Resour. Policy* **2021**, *74*, 102254. [CrossRef]
3. Lin, Y.; Yang, H.; Ma, L.; Li, Z.; Ni, W. Low-Carbon Development for the Iron and Steel Industry in China and the World: Status Quo, Future Vision, and Key Actions. *Sustainability* **2021**, *13*, 12548. [CrossRef]
4. Fan, Z.; Friedmann, S.J. Low-carbon production of iron and steel: Technology options, economic assessment, and policy. *Joule* **2021**, *5*, 829–862. [CrossRef]
5. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
6. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Part III 18; Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
7. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
8. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
9. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
10. Wang, H.; Zhu, Y.; Adam, H.; Yuille, A.; Chen, L.C. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5463–5474.
11. Zhang, H.; Li, F.; Xu, H.; Huang, S.; Liu, S.; Ni, L.M.; Zhang, L. MP-Former: Mask-piloted transformer for image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 18074–18083.
12. Jain, J.; Li, J.; Chiu, M.T.; Hassani, A.; Orlov, N.; Shi, H. Oneformer: One transformer to rule universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 2989–2998.
13. Tragakis, A.; Kaul, C.; Murray-Smith, R.; Husmeier, D. The fully convolutional transformer for medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Vancouver, BC, Canada, 18–22 June 2023; pp. 3660–3669.
14. Fujieda, S.; Takayama, K.; Hachisuka, T. Wavelet Convolutional Neural Networks. *arXiv* **2018**, arXiv:cs.CV/1805.08620. Available online: http://arxiv.org/abs/1805.08620 (accessed on 1 January 2024).
15. Liu, P.; Zhang, H.; Zhang, K.; Lin, L.; Zuo, W. Multi-level Wavelet-CNN for Image Restoration. *arXiv* **2018**, arXiv:cs.CV/1805.07071. Available online: http://arxiv.org/abs/1805.07071 (accessed on 1 January 2024).

16. Wu, T.; Li, W.; Jia, S.; Dong, Y.; Zeng, T. Deep Multi-Level Wavelet-CNN Denoiser Prior for Restoring Blurred Image with Cauchy Noise. *IEEE Signal Process. Lett.* **2020**, *27*, 1635–1639. [CrossRef]
17. Huang, H.; He, R.; Sun, Z.; Tan, T. Wavelet-SRNet: A wavelet-based CNN for multi-scale face super resolution. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1698–1706. [CrossRef]
18. Ma, H.; Liu, D.; Xiong, R.; Wu, F. iWave: CNN-Based Wavelet-Like Transform for Image Compression. *IEEE Trans. Multimed.* **2020**, *22*, 1667–1679. [CrossRef]
19. Kim, C.W.; Kim, H.G. Study on automated scrap-sorting by an image processing technology. *Adv. Mater. Res.* **2007**, *26*, 453–456. [CrossRef]
20. Koyanaka, S.; Kobayashi, K. Automatic sorting of lightweight metal scrap by sensing apparent density and three-dimensional shape. *Resour. Conserv. Recycl.* **2010**, *54*, 571–578. [CrossRef]
21. Wieczorek, T.; Pilarczyk, M. Classification of steel scrap in the EAF process using image analysis methods. *Arch. Metall. Mater.* **2008**, *53*, 613–617.
22. Xu, G.; Li, M.; Xu, J. Application of machine learning in automatic grading of deep drawing steel quality. *J. Eng. Sci.* **2022**, *44*, 1062–1071.
23. Duan, S. Recognition Classification and Statistics of Scrap Steel Based on Optical Image YOLO Algorithm. Master's Thesis, Dalian University of Technology, Dalian, China, 2021.
24. Xu, W.; Xiao, P.; Zhu, L.; Zhang, Y.; Chang, J.; Zhu, R.; Xu, Y. Classification and rating of steel scrap using deep learning. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106241. [CrossRef]
25. Sun, L. Automatic rating of scrap steel based on neural network. *Chin. Informatiz.* **2021**, 49–50.
26. *GB/T 4223-2017*; Iron and Steel Scraps. China National Standards: Beijing, China, 2017.
27. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
28. Zhang, C.; Kim, J. Modeling long-and short-term temporal context for video object detection. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 71–75.
29. Liang, X.; Shen, X.; Xiang, D.; Feng, J.; Lin, L.; Yan, S. Semantic object parsing with local-global long short-term memory. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3185–3193.
30. Bracewell, R.; Kahn, P.B. The Fourier transform and its applications. *Am. J. Phys.* **1966**, *34*, 712. [CrossRef]
31. Geng, Z.; Guo, M.H.; Chen, H.; Li, X.; Wei, K.; Lin, Z. Is attention better than matrix decomposition? *arXiv* **2021**, arXiv:2109.04553.
32. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
33. Lee, Y.; Kim, J.; Willette, J.; Hwang, S.J. Mpvit: Multi-path vision transformer for dense prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7287–7296.
34. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
35. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
36. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
37. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.