IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 39, No. 2, April, 2022, pp. 701-710
Journal homepage: http://iieta.org/journals/ts

Automatic Classification and Identification of Road Garbage Images and Evaluation of Environmental Health Based on UNet++

Wei Chen[1], Xuan Zheng[2], Haijun Zhou[3], Zhe Li[4,5*]

[1] Logistics Service Department, Wenzhou Business College, Wenzhou 325035, China
[2] Information Service Center, Wenzhou Business College, Wenzhou 325035, China
[3] Zhejiang College of Security Technology, Wenzhou 325016, China
[4] Institute of Economics and Management, Hubei Engineering University, Xiaogan 432000, China
[5] Research Center of Hubei Micro & Small Enterprise, Xiaogan 432000, China

Corresponding Author Email: lizhe_lz@hbeu.edu.cn

https://doi.org/10.18280/ts.390234

Received: 3 January 2022
Accepted: 5 March 2022

Keywords:
semantic segmentation, UNet++, garbage classification, eye tracking analyzer, attention mechanism, evaluation method of environmental health

ABSTRACT

As Covid-19 plagues the world, a clean environment helps to control the factors and risks that threaten health, and curb the spread of the epidemic. However, the quality evaluation of environmental health faces some problems and challenges in actual management and practice. Firstly, the classification, identification, and quantification of road garbage are mainly done manually, because of the diversity of road garbage, as well as their sharp differences in geometry, color, and texture. Secondly, it is labor-intensive to manually manage the large operation areas on the wide urban roads. Thirdly, the accuracy of statistical indices is affected by the time-varying road environment, making the quality evaluation of environmental health untimely and inaccurate. To solve these problems, this paper proposes an intelligent image classification and evaluation method for urban environmental health. Specifically, an environmental garbage recognition and semantic segmentation approach was designed based on UNet++, and combined with the vehicle-mounted machine vision system to automatically identify the typical targets among the road waste control indices. Next, an image attention quantitative evaluation method was developed based on the eye tracking analyzer, and the quantified attention was fused with the statistical features for road garbage classification, forming an attention-based evaluation method for environmental quality. The proposed approach supports the automatic recognition and semantic segmentation of the garbage on urban roads, and realizes the identification of complex targets in different scenes through transfer learning. In addition, the attention-based evaluation method for environmental quality provides environmental management departments with visual basis for quantitative decision-making.

1. INTRODUCTION

As the world becomes more and more urbanized, municipal garbage stands out as the most important by-product of urban life. The growth rate of municipal garbage is even faster than that of urbanization. According to the latest research report, 3.4 billion tons of waste will be generated annually by 2050, a substantial increase from the current amount of 2.01 billion tons [1]. Road garbage, a kind of urban garbage, occupies a huge area of land, affects the urban landscape, and even causes irreversible damage to the environment. Therefore, more and more attention has been paid to road garbage [2]. As the mainstay of urban garbage, road garbage seriously tarnishes the urban image, and hinders the long-term development of cities. It also directly shapes the subjective feelings of city dwellers [3]. As a result, many countries consider garbage and debris as primary indices of urban road environment, when they formulate management and evaluation measures for urban environmental health [4]. Nevertheless, the quality evaluation of environmental health faces the following problems and challenges in actual management and practice:

Firstly, the classification, identification, and quantification of road garbage are mainly done manually, because of the diversity of road garbage, as well as their sharp differences in geometry, color, and texture. Secondly, it is labor-intensive to manually manage the large operation areas on the wide urban roads. Thirdly, the accuracy of statistical indices is affected by the time-varying road environment, making the quality evaluation of environmental health untimely and inaccurate. In fact, the current evaluation methods for environmental health quality rarely consider the subjective feeling and sensitive of different people to different types of garbage.

With the continuous development of artificial intelligence (AI), automation and intellectualization can be adopted to reduce the dependence on manual operations, providing an effective means to overcome labor shortage, reduce labor intensity, and improve work efficiency. Hence, an important, feasible way to solve the said problems and challenges is to detect road garbage, and automatically summarize the indices of environmental quality through image recognition, coupled with better evaluation criteria.

In recent years, intelligent technologies like deep learning have been widely applied in multiple fields, such as natural language processing, computer vision, and semantic comprehension [5], resulting in major breakthroughs in these aspects [6, 7]. On many public datasets, deep learning-based methods have demonstrated high accuracy [8]. The high practicability of deep learning, and its good performance in computer vision pique the interests of researchers and practitioners. It is interesting to consider the application of deep learning in garbage recognition.

Some of the traditional studies on garbage classification are reviewed below. Using k-means clustering (KMC) algorithm, Niska et al. [9] identified and classified building garbage by color features. Ge et al. [10] developed a combinatory algorithm of support vector machine (SVM), and applied it to recognize marine garbage images. Miraliakbari et al. [11] identified the cracks in road images based on different texture features. Zalama et al. [12] combined Gabor filter and AdaBoost classifier to improve the recognition of road cracks. However, the above detection approaches are not robust, and relatively complex [13]. Besides, it is difficult to achieve a high recognition rate by the traditional segmentation strategies, for the road garbage is smaller, coarser, and less convex/concave than the asphalt pavement.

In actual applications (e.g., the evaluation index system of the environment), the quantitative indices include the area affected by garbage, in addition to the type of garbage [14]. Therefore, the contour of target garbage should be extracted from the original image through semantic segmentation, before recognizing garbage type and quantity through classification. Mittal et al. [15] detected garbage and segmented the garbage areas with the fully convolutional network (FCN), and achieved a test accuracy of 87.69%. But the FCN made many mistakes in garbage identification. Wei et al. [16] constructed a garbage classification model based on the faster region-based convolutional neural network (Faster-RCNN), which boasts high reliability and good real-time performance. However, the detection accuracy of the network plunged, when the samples differ significantly from the training model in scale [17]. Chen et al. [18] created DeepLab based on the FCN, and introduced dilated convolution to solve the information loss, which arises from the reduction in the resolution of the feature map. Ronneberger et al. [19]proposed the UNet, which relies on a few data to realize end-to-end training. To enhance the network ability of global information acquisition, Zhao et al. [7] put forward the pyramid scene parsing network (PSPNet). All these approaches enhance the accuracy of semantic segmentation of road garbage.

By virtue of the deep learning technology of AI, this paper designs a deep neural network (DNN) and a CNN for segmenting and classifying the soils, gravels, and leaves of the road sweeper truck. The proposed tools reduce the time and labor costs of road garbage recognition and quantification. Next, the attention of evaluators to the garbage in images was acquired by an eye tracking analyzer, and this intrinsic subjective cognition was quantified by the attention quantification model, producing a more accurate evaluation model. Our model fully considers the effects of human factors, and improves the scientific level of environmental quality evaluation.

2. RESEARCH FRAMEWORK

Figure 1 shows the overall research framework. There are four major parts of the system: Firstly, the vehicle-mounted machine vision system collects images and videos of the environment, providing the data basis for subsequent analysis. Secondly, the UNet++-based garbage classification and semantic segmentation model recognizes the type and quantity (scale) of garbage in the environment, laying the quantitative basis for environmental quality evaluation. Thirdly, the eye tracking analyzer collects and quantifies the attention of personnel, according to different scenes and types of garbage, in order to build an attention-based weight model for garbage classification. Fourthly, the scores of environmental health quality are rated according to the quantitative values in the second part and the garbage weights in the third part.

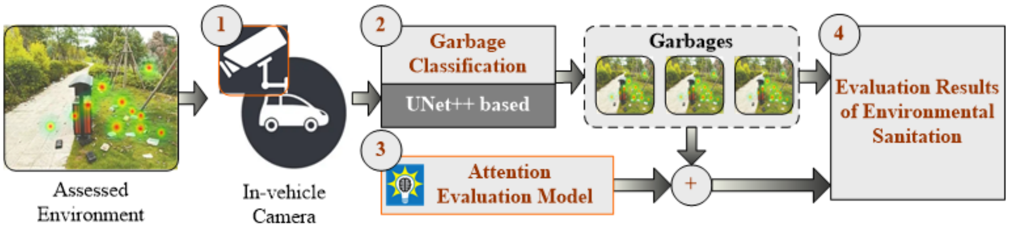


Figure 1. Overall research framework

2.1 UNet++-based urban road garbage classification and semantic segmentation model (baseline)

The ordinary neural networks generally utilize low-level features of images, which constrain the ability of classification and semantic segmentation. In the urban environment, the unobvious small-scale features, such as the road textures, sands, or soils, would suppress the effect of the classifier. To solve the problem, this paper proposes a road garbage recognition approach based on the UNet++, a new and powerful architecture for image segmentation. In essence, the architecture is a deeply supervised encoder-decoder network. The subnets of the encoder and decoder are linked up by a series of nested skip connections. The redesigned skip connections enable the optimizer to learn the relevant tasks more easily. The information gap between the feature maps of the encoder and decoder is reduced, such that the feature mapping between the two modules has semantic similarity.

2.2 Construction of UNet++ model

The UNet++ differs from the classic UNet [19] in the following aspects: An up-sampling is added to each down-sampling, followed by a skip connection. On the skip connection, dense convolution blocks are applied to adjust the information gap between feature maps on different layers. Hence, the UNet++ has the same up-sampling and down-sampling modules as the UNet. Like the FCN, the UNet is a tool based on the encoder-decoder structure. The defining feature of the UNet is the absolute symmetry. The network is composed of a contraction path and an expansion path. The former is responsible for acquiring the contextual information, while the latter is responsible for precise positioning. The two paths are symmetric to each other. Figures 2 and 3 show the topology and architecture of the UNet, respectively. The UNet begins with five down-sampling layers, followed by the corresponding up-sampling layers. After each up-sampling, the features of the previous five convolutional layers are linked up via a skip connection. In the final output layer, a 1×1 convolutional kernel, and an activation function, i.e., the rectified linear unit (ReLU), are added for deep supervision. Finally, the output layer provides the detected pixels in the original road garbage image.

In Figure 3, each bounding box corresponds to a multi-channel feature map. The top of the box is the number of channels. The lower left of the box is the map size. Each copied feature map is illustrated as a hollow bounding box. Different operations are represented by arrows, including convolutional kernel, copy and crop, max pooling, and up-convolution. It can be observed that the UNet focuses on the deep information of the original image. Up-sampling is not carried out until the $X^{4,0}$-th layer. Besides, there is a huge semantic gap between the encoder and decoder. Nevertheless, the feature maps contain refined lower-layer structures, and offer many visual information of the original image, including boundaries and colors. These information plays an important role in the recognition of road garbage. Hence, the authors decided to add up-sampling to the $X^{1,0}$-th, $X^{2,0}$-th, $X^{3,0}$-th, and $X^{4,0}$-th layers, i.e., apply the improved UNet (UNet++) to the identification of road garbage.

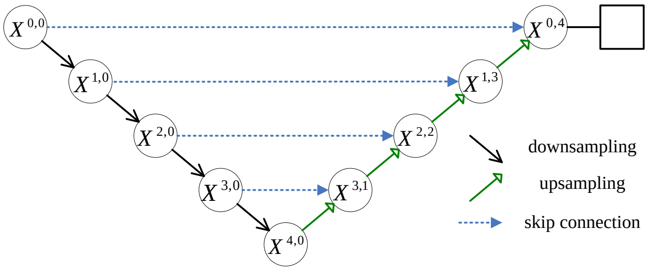


Figure 2. Topology of UNet

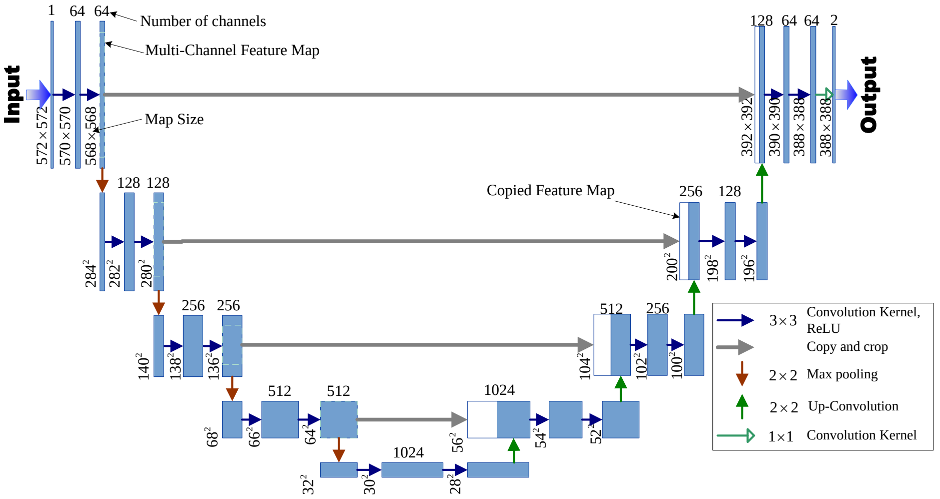


Figure 3. Architecture of UNet (an example with the lowest resolution of 32×32 pixels)

Figure 4 shows the overall structure of UNet++. It can be observed that the network starts from an encoder sub-network, or the backbone of that subnetwork. The main difference between UNet++ and UNet lies in the incorporation of up-sampling and skip connection. Each skip connection links up the two sub-networks, and the deep supervision parts of the map.

2.2.1 Redesign of skip connections

The redesign of skip connections changes the connectivity between the encoder and decoder in the UNet++. In the UNet, the decoder directly obtains the feature map of the codes. In the UNet++, the feature map provided by the decoder passes through a dense convolution block, in which the number of convolutional layers depends on the level of pyramid in the network. As shown in Figure 4, there are three layers of dense convolution blocks along the skip connection between nodes $X^{0,0}$ and $X^{1,3}$. Each convolution layer is preceded by a connection layer, which fuses the output of the convolutional layer preceding that dense block with the output of the sparse block of the corresponding up-sampling layer. The fusion ensures that the feature mappings of the encoder and decoder have similar information, which improves the efficiency of the optimizer.

The skip connection can be expressed mathematically. Let $x^{i,j}$ be the output of node $X^{i,j}$; i be the down-sampling layer along the encoding direction; j be the convolutional layer of the dense block along the skip connection. Then, the value of $x^{i,j}$ can be calculated by:

$$x^{i,j} = \begin{cases} \mathcal{H}\left(x^{i-1,j}\right) & j = 0 \\ \mathcal{H}\left(\left[x^{i,k}\right]_{k=0}^{j-1}, \mathcal{U}\left(x^{i+1,j-1}\right)\right) & j > 0 \end{cases} \tag{1}$$

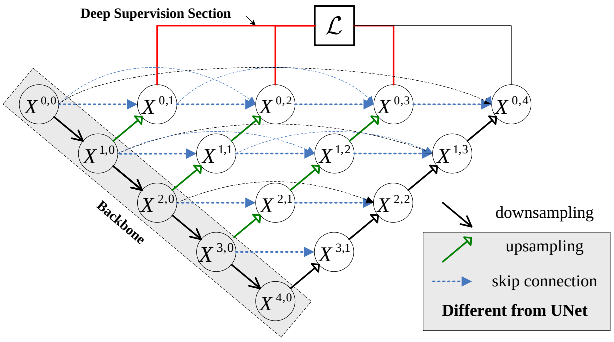


Figure 4. Overall structure of UNet++

where, function $\mathscr{H}(\bullet)$ is the convolution of the feature map (each convolutional layer is followed closely with an activation function); $\mathcal{U}(\bullet)$ is the up-sampling operation; $[\bullet]$ is the connection layer; $j = 0$ is the node that receives an input from the previous layer; $j = 1$ is the node that receives two inputs, both of which come from two continuous layers in the encoder sub-network; $j > 1$ is the node that receives j+1 inputs, in which j inputs are outputted by the nodes preceding the skip connection consistent with the node, and the remaining 1 input is obtained by up-sampling in the skip connection lower than that node. Along each skip connection, a dense connection block is utilized to preserve the previous feature maps, and enter them into the right node. Figure 5 provides an example of how feature maps pass through the top skip connection of the UNet++, and drives the value of $x^{i,j}$ by formula (1).

2.2.2 Deep supervision

Deep supervision is adopted in the UNet++ [20], allowing the model to operate in two modes: (1) The precision mode: the output is the average of all segmentation branches; (2) The rapid mode: The final segmentation map only comes from one of the segmentation branches. The selection of the branch determines the pruning degree and speed gain of the model. Figure 6 explains the variation of architecture complexity with the selection of different branches under the rapid mode, taking the UNet++L1 network with three layers pruned as an example.

Based on the nested skip connections, deep supervision was added to $\left\{X^{0,j}, j \in \{1,2,3,4\}\right\}$, so as to form a full resolution feature map on multiple semantic levels. For instance, the combination of binary cross entropy and the Dice coefficient could be added to each of the four semantic layers as the loss function, thereby formulating the full resolution feature map:

$$\mathcal{L}\left(Y,\hat{Y}\right) = -\frac{1}{N}\sum_{b=1}^{N}\left(\frac{1}{2}Y_b \log \hat{Y}_b + \frac{2Y_b\hat{Y}_b}{Y_b + \hat{Y}_b}\right) \tag{2}$$

where, $\hat{Y}_b$ and Y_b are the flatten predicted probability and the flatten ground truth of the b-th image, respectively; N is the batch size.

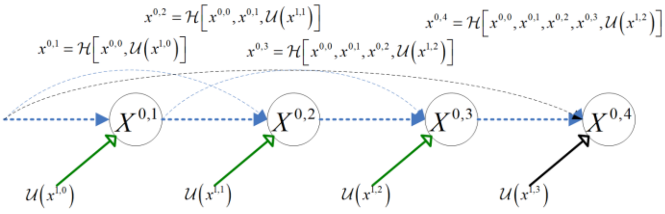

$x^{0,2} = \mathcal{H}\left[x^{0,0}, x^{0,1}, \mathcal{U}\left(x^{1,1}\right)\right]$
$x^{0,4} = \mathcal{H}\left[x^{0,0}, x^{0,1}, x^{0,2}, x^{0,3}, \mathcal{U}\left(x^{1,2}\right)\right]$
$x^{0,1} = \mathcal{H}\left[x^{0,0}, \mathcal{U}\left(x^{1,0}\right)\right]$
$x^{0,3} = \mathcal{H}\left[x^{0,0}, x^{0,1}, x^{0,2}, \mathcal{U}\left(x^{1,2}\right)\right]$
$X^{0,1}$
$X^{0,2}$
$X^{0,3}$
$X^{0,4}$
$\mathcal{U}\left(x^{1,0}\right)$
$\mathcal{U}\left(x^{1,1}\right)$
$\mathcal{U}\left(x^{1,2}\right)$
$\mathcal{U}\left(x^{1,3}\right)$

Figure 5. An example of the feature map processing through the top skip connection

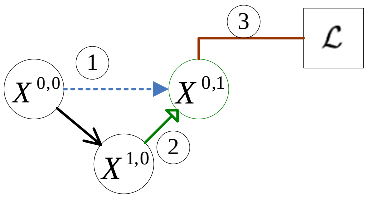


Figure 6. Structure of UNet++L1

Table 1. Skip connection parameters

Original layer	Target layer	Original layer	Target layer
$X^{0,0}$	$X^{0,1}$, $X^{0,2}$, $X^{0,3}$, $X^{0,4}$	$X^{1,1}$	$X^{1,2}$, $X^{1,3}$
$X^{0,1}$	$X^{0,2}$, $X^{0,3}$, $X^{0,4}$	$X^{1,2}$	$X^{1,3}$
$X^{0,2}$	$X^{0,3}$, $X^{0,4}$	$X^{2,0}$	$X^{2,1}$, $X^{2,2}$
$X^{0,3}$	$X^{0,4}$	$X^{2,1}$	$X^{2,2}$
$X^{1,0}$	$X^{1,1}$, $X^{1,2}$, $X^{1,3}$	$X^{3,0}$	$X^{3,1}$

Note: $X^{0,1}$, $X^{0,2}$, $X^{0,3}$, and $X^{0,4}$ stand for deep supervision.

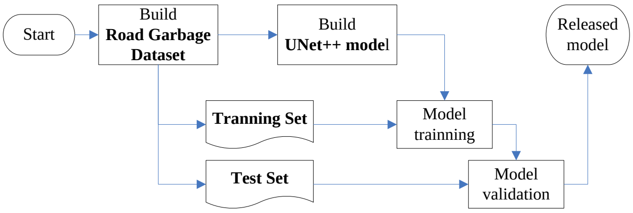


Figure 7. Flow of UNet++-based road garbage identification

Following the illustration in Figure 6, the UNet++ was improved in three aspects: (1) the convolutional layer on the skip connection 1 reduces the information gap between the feature maps corresponding to the encoder and decoder; (2) the skip connection 2 mitigates the gradient flow; (3) the skip connection 3 represents deep supervision, which improves the UNet++ through pruning. In this way, this paper constructs the structure of the UNet++, which share the same structure of encoder and decoder as the UNet. Table 1 shows the skip connection parameters of the UNet++.

The proposed UNet++ was trained and tested on a self-designed road garbage dataset.

2.3 Training strategy

The UNet++-based road garbage identification method is implemented in four steps: data preprocessing, UNet++ modeling, model training, and model testing (Figure 7).

Step 1. The original data are preprocessed, and the dataset is divided into a training set, a verification set, and a test set.

Step 2. The UNet++ model is constructed, and the relevant parameters are initialized.

Step 3. The UNet++ model is trained by the training set.

Step 4. The road garbage recognition accuracy of the UNet++ model is tested on the test set.

3. ATTENTION-BASED EVALUATION MODEL

3.1 Attention quantification mechanism based on eye tracking analyzer

This paper designs an attention quantification model, which accurately evaluates the attention of personnel through four steps. Figure 8 illustrates the framework of the attention quantification mechanism based on eye tracking analyzer.

Step 1. Setting up test procedure

The same test procedure is set up for each experiment and tester.

Step 2. Adding stimuli

The test images containing garbage are introduced. The videos are converted into image frames.

Step 3. Setting stimuli properties

The stimuli properties are configured, including the size, brightness equalization, stimulation time, and garbage type of test images.

Step 4. Preview

The experiments are previewed.

Step 5. Adjusting the order of stimuli

The order of stimuli is adjusted based on the preview results. Normally, the test data of different types are arranged alternatively to avoid the influence of empiricism. Otherwise, the samples will become less sensitive, causing significant test errors.

Step 6. Instrument calibration

Before each experiment, every tester must calibrate the instrument to minimize the test deviations.

Step 7. Attention quantification

The attention is quantified based on the discrete eye movement foci with sequential labels, according to the rules.

Figure 9 shows the experimental environment of eye tracking analyzer, which mainly includes the eye tracker, the stimuli (test images), and focus data (each focus is converted into a heat map of attention).

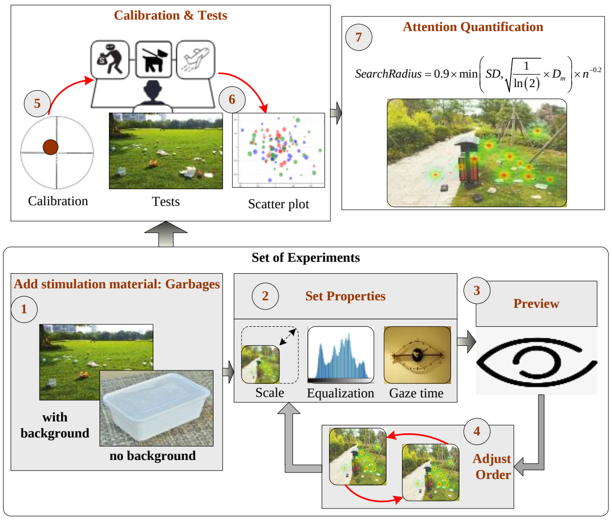


Figure 8. Attention quantification mechanism based on eye tracking analyzer

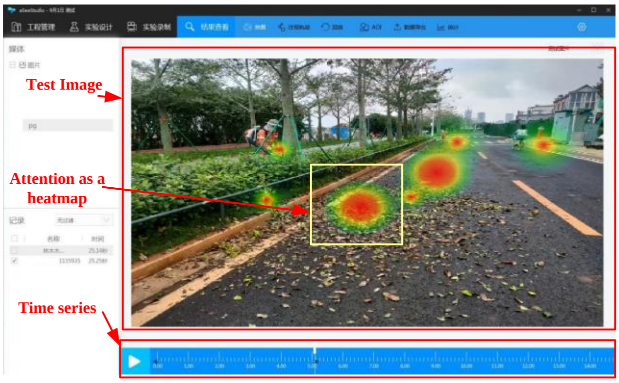
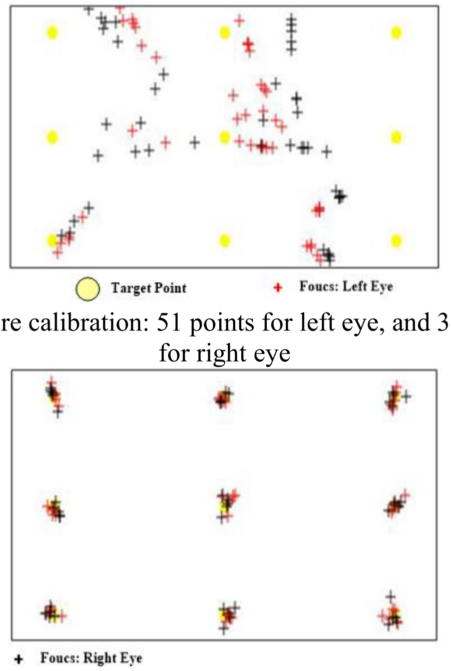


Figure 9. Experimental environment of eye tracking analyzer

3.2 Eye tracker calibration

The possible changes in instrument position or subject pose may affect the accuracy of eye movement data. Thus, the eye tracker must be calibrated prior to each experiment. This paper measures the calibration quality by graphs and calibration scores (Figure 10). Each subject was asked to stare at a fixed yellow target point on the calibration image. The focus of the eyes (red point for the left eye and black point for the right eye) was presented as scatter points of attention. The calibration quality was judged by the aggregation of scatter points, and the coincidence with the target point. To ensure the quality of eye movement data, the calibration score must be greater than 85.

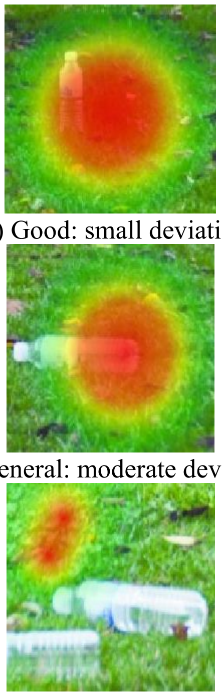


(a) Before calibration: 51 points for left eye, and 39 points for right eye

(b) After calibration: 92 points for left eye, and 89 points for right eye

Figure 10. Instrument calibration before experiment

3.3 Deviation processing



(a) Good: small deviation

(b) General: moderate deviation

(c) Poor: large deviation

Figure 11. Deviations generated in actual test

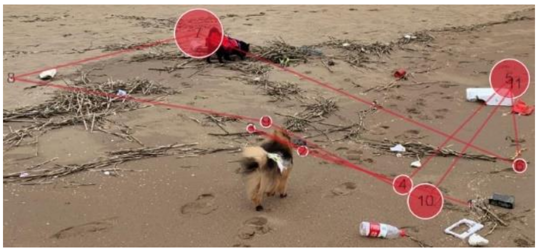


Figure 12. Fixation trajectory

Under the influence of multiple factors, the focus data obtained in the actual test cannot completely overlap the target, as shown in Figure 11.

The main reasons for the deviations are as follows: the attention interference caused by other objects in the test scene; changes in the position of the subject's eyes during the test; deviations produced by the subject's habits.

In addition, during the test, the subject's eyes are not fixed at a position, but move along a trajectory composed of a series of discrete points (Figure 12). When the discrete data are converted into attention, the data beyond the fixation target must be filtered out. If the filtering is incomplete, the attention quantification may be incorrect.

3.4 Attention quantification

As shown in Figure 12, the original fixation trajectory is made up of discrete points. Attention quantification is needed to apply the trajectory to quantitative evaluation of the attention of different types of garbage.

Referring to the literature, the common practice is to convert the discrete points into a heat map. The point density is visualized as a heat map, using the density function. The spatial position and basic features of spatial distribution are highlighted, allowing people to perceive point density independent of the scale factor. There are two classic density functions, namely, the kernel density analysis, and the point kernel density analysis.

The kernel density analysis computes the density of elements in the neighborhood. By interpolating discrete point data, different weights are assigned to the points within the search range. Large weights are given to the points or lines close to the search center, and small weights are given to those far from the center. The calculation results of this tool are smoothly distributed.

The point kernel density analysis solves the point density around each output grid cell. Each point is covered with a smooth surface. The surface value peaks at the position of the point, and gradually decreases as the distance from the point increases. The surface value is zero, when the distance from the point is equal to the search radius. Only circular neighborhoods are allowed [21].

The eye movement trajectory, as the dynamic psychological activity of the subjects, cannot be simply removed. Therefore, this paper adopts the search radius (bandwidth) algorithm. The main process of the algorithm is as follows:

Step 1. Calculate the average center of the input points.

Step 2. Calculate the distance to the (weighted) average center of all points.

Step 3. Calculate the (weighted) median of these distances.

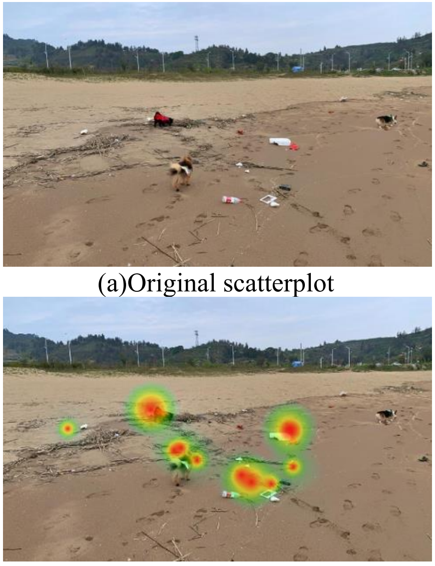
Step 4. Calculate the (weighted) standard distance.

Step 5. Calculate the bandwidth using the following formula:

$$SearchRadius = 0.9 \times \min\left(SD, \sqrt{\frac{1}{\ln(2)}} \times D_m \right) \times n^{-0.2} \tag{3}$$

where, $min(\bullet)$ is the smaller one between the results of the two functions.

As shown in Figure 13, the original scatterplot of the attention is converted into a heat map of attention through the above calculations.



(a)Original scatterplot

(b) Calculated heat map

Figure 13. Attention quantification

4. EXPERIMENTS AND RESULTS ANALYSIS

4.1 Datasets

There are two groups of samples for our experiments. One group was collected and sorted from the Internet, and the other was shot by our research team. The typical samples were acquired from the urban environment, including roads, parks, river / sea beaches, etc. By the complexity of the scene, the samples were divided into three classes, namely, roads, green fields, and hybrid scenes (Figure 14). The test scenes apart from roads intend to enhance the generalization ability of our model: the other interferences make the simulation environment more realistic.

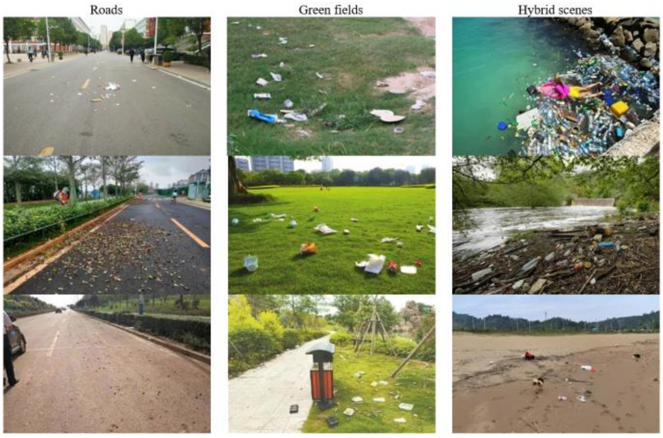
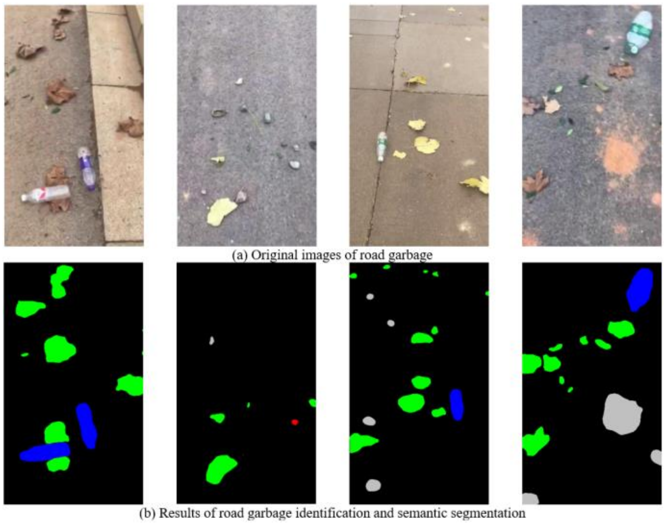


Figure 14. Three test scenes

The total sample size was 8,700, including 4,200 road images, 2,100 green field images, and 4,300 hybrid scene images.

To further enhance the garbage classification and identification ability, 14,000 region-of-interest (ROI) images of garbage were collected and sorted out. Figure 15 shows the samples of the garbage "plastic bowls".

Figure 15. Garbage dataset with labels


(a) Original images of road garbage

(b) Results of road garbage identification and semantic segmentation

Figure 16. Visualization of training results for UNet++ model

4.2 Test results on recognition effect

The Adam optimizer was adopted, the learning rate was initialized as 0.001 for the training, the number of iterations was set to 50, and the training set, verification set, and test set were prepared by the ratio of 6:1:3. The original images were enhanced 10 times. The trained UNET++ model was adopted to recognize the road garbage in the test set. Figure 16 visualizes some test results. The black, green, blue, gray, and red parts in the figure represent the background, leaves, bottles, sand piles, and stone blocks detected in the experiment.

The garbage recognition effect of the UNet++ model was measured by pixel accuracy (PA), mean pixel accuracy (MPA), and mean intersection over union (MIoU). The scores of the model on the three metrics were 0.972, 0.72, and 0.731, respectively. The recognition effect of the model on six different kinds of garbage was analyzed independently, against the metric of PA. Table 2 shows the recognition results of the UNet++ model on different types of garbage.

Table 2. Recognition results of the UNet++ model on different types of garbage

Class	Result		
Pericarp	0.792	Sandstone	0.889
Scraps and plastic film	0.804	Water stains	0.923
Cigarette butt	0.686	Other	

As shown in Table 2, the UNet++ model achieved an overall good recognition rate of road garbage. The recognition rate was relatively high on pericarp, scarps and plastic film, sandstone, and water stains, and relatively poor on small targets like cigarette butt.

4.3 Test results on attention mechanism

The test instrument is aSee Pro (7invensun, Beijing, China), with a sampling rate of 250Hz and an accuracy of <0.5°. The attention experiment was designed based on the aSee Studio software. The attention experiment on each sample follows the steps in Figure 8. Figure 17 visualizes the heat maps of attention for environmental garbage in different scenes.

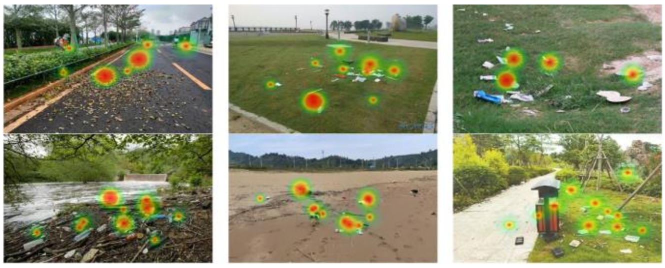


Figure 17. Visualization of the heat maps of attention for environmental garbage in different scenes

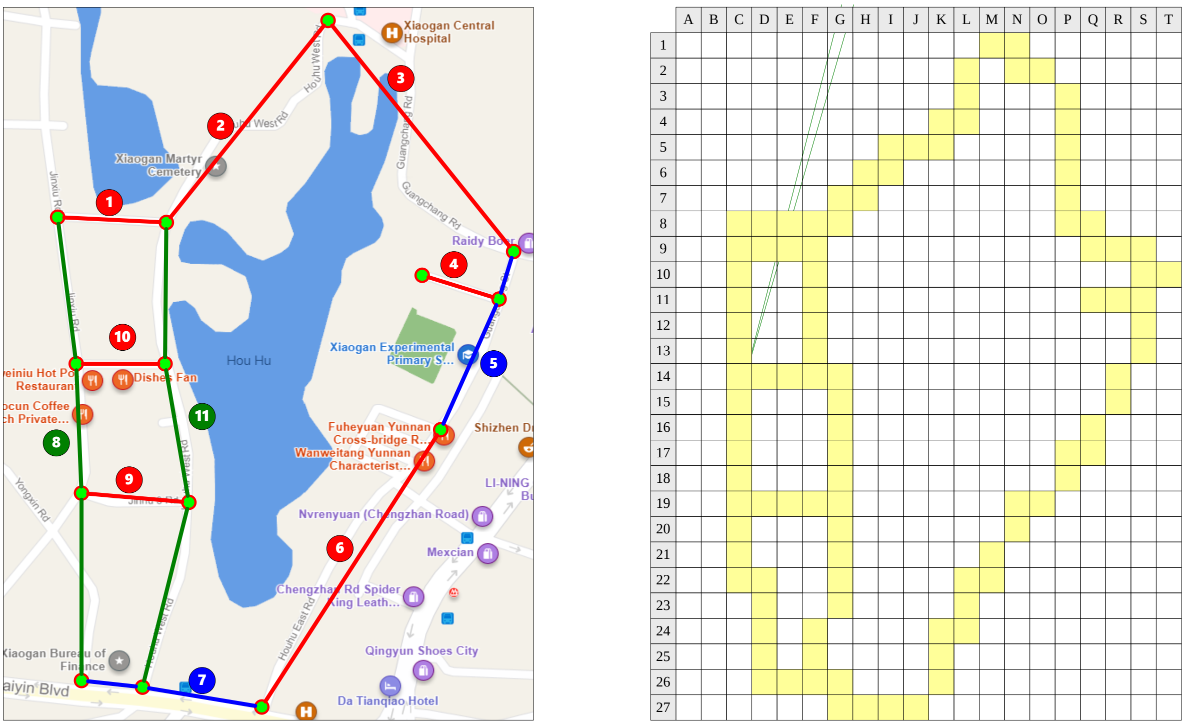
Table 3. Garbage classes and attention weights after normalization

Garbage class	Mean attention	Garbage class	Mean attention
Pericarp	0.171	Sandstone	0.101
Scraps and plastic film	0.347	Water stains	0.231
Cigarette butt	0.052	Other	0.15

Table 3 shows the learning results of attention weights. Among the six types of targets, the most prominent scraps and plastic film achieved the highest attention score, owing to their frequent appearance and eye-catching colors. No wonder they are called the white garbage. The water stains also attracted much attention, for their large size and prominent features. By contrast, cigarette butt attracted the least attention. On the one hand, this type of garbage is very small, and thus unlikely to cause aversion from afar. On the other hand, cigarette butt only appears in a small range, because many countries have adopted non-smoking measures.

4.4 Comparative analysis of environmental health indices based on attention mechanism

Multiple samples were collected from a test area in the city of the authors. The collected samples are included in the sample set in Figure 14. As shown in Figure 18(a), the test area includes 11 roads (segments), with a total length of 3,100m. The size of the test area is about 0.22km^2. The test area is strongly representative, as it involves multiple typical urban scenes, such as schools, hospitals, parks, public squares, restaurants, hotels, and residential neighborhoods. As shown in Figure 18, the density analysis of the test area was implemented, using a rasterized map. The grid size was 50×50m. The road garbage was counted based on the samples collected on December 1st, 2021. The results are displayed in Table 4.



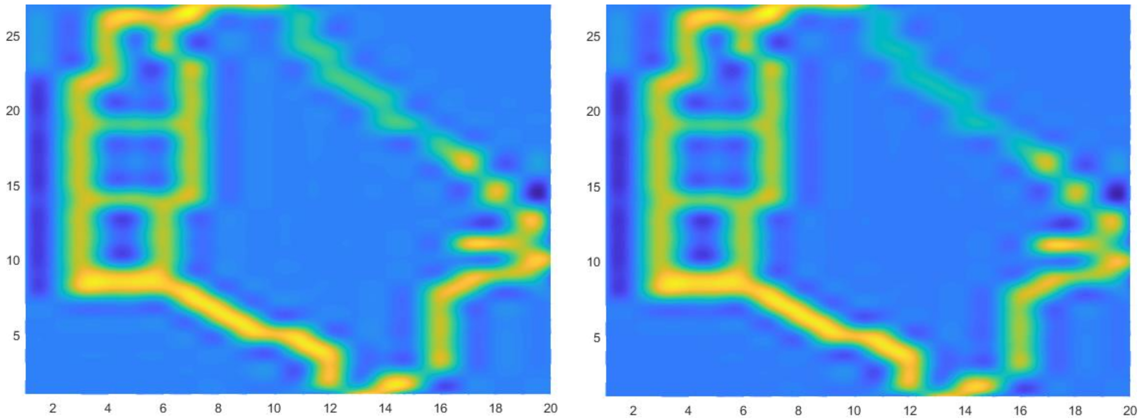
(a) Map of the test area

(b) Rasterized map

Figure 18. The test area, including 11 roads

Table 4. Statistics on each class of garbage in the test area

Street No.	Quantity of garbage						Statistical score		AM score	
	1	2	3	4	5	6	Level	AVG	Level	AVG
1	2	3	1	2	0	3	1	4.67	1	4.13
2	0	1	1	1	0	2	1	4.83	1	4.26
3	4	5	4	0	1	6	2	4.00	2	3.27
4	1	1	2	0	0	2	1	4.83	1	4.26
5	5	7	1	0	2	3	2	3.83	2	3.12
6	9	11	6	2	3	9	4	2.33	4	1.49
7	0	2	1	0	0	2	1	4.83	1	4.26
8	1	4	1	7	1	4	3	4.00	3	3.47
9	2	0	2	12	4	11	4	2.83	4	2.54
10	3	5	7	1	3	7	3	3.50	3	2.91
11	1	3	2	5	2	7	2	3.83	2	3.35



(a) Statistical method

(b) Attention mechanism

Figure 19. Comparison of environmental health quality scores: heat maps

As shown in Figure 19, the scores obtained by the proposed evaluation method were close to human perception.

5. CONCLUSIONS

This paper explores the practical problems in the management of urban environmental health, and proposes to automatically identify and classify road garbage through deep learning, aiming to effectively reduce manual labor intensity, and enhance working efficiency. By quantifying the subjective feelings and the sensitivity to garbage in different scenes, the authors established an environmental quality evaluation model based on the attention mechanism. The model is in line with the idea of people-oriented management, compared with the statistical evaluation approach.

Nonetheless, the urban health environment is a thorny management issue under complex backgrounds. The changeable environment and scenes, coupled with the diversity of garbage, make it very challenging to develop garbage identification models with a strong generalization ability. Take our case study for example. The small sample recognition rate remained low in the hybrid scenes. In addition, it takes a long cycle to apply the attention-based evaluation method to the management practice. It is necessary to pilot and evaluate our approach in a greater range. The new evaluation and management system based on our approach needs serious deliberations.

ACKNOWLEDGMENT

This project was funded by the Humanities and Social Science Research Project of Ministry of Education of China (Grant No.: 20YJCZH081), and Scientific Research Project of Education Department of Hubei Province (Grant No.: D20212701, D20202701, B2019389).

REFERENCES

[1] Kaza, S., Yao, L.C., Bhadatata, P., Van Woerden, F. (2018). What a waste 2.0. World Bank Publications. https://openknowledge.worldbank.org/handle/10986/17 388 License: CC BY 3.0 IGO.

[2] Appiah, J.K., Berko-Boateng, V.N., Tagbor, T.A. (2017). Use of waste plastic materials for road construction in ghana. Case Studies in Construction Materials, 6(C): 1-7. https://doi.org/ 10.1016/j.cscm.2016.11.001

[3] Samuelsson, K., Giusti, M., Peterson, G.D., Legeby, A., Brandt, S.A., Barthel, S. (2018). Impact of environment on peoples everyday experiences in stockholm article in fo. Landscape and Urban Planning, 171: 7-17. https://doi.org/10.1016/j.landurbplan.2017.11.009

[4] Irvine, K.N., Fuller, R.A., Devine-Wright, P., et al. (2010). Ecological and psychological value of urban green space. Dimensions of the Sustainable City, pp. 215-237. https://doi.org/10.1007/978-1-4020-8647-2_10

[5] Lu, H., Zhang, Q. (2016). Applications of deep convolutional neural network in computer vision. Journal of Data Acquisition and Processing, 31(1): 1-17. https://doi.org/10.16337/j.1004-9037.2016.01.001

[6] Long, J., Shelhamer, E., Darrell, T. (2016). Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 3431-3440. https://doi.org/10.1109/cvpr.2015.7298965

[7] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2016). Pyramid scene parsing network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 6230-6239. https://doi.org/10.1109/CVPR.2017.660

[8] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Zitnick, C. L. (2014). Microsoft coco: common objects in context. 13th European Conference, Zurich, Switzerland, pp. 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[9] Niska, H., Serkkola, A. (2018). Data analytics approach to create waste generation profiles for waste management and collection. Waste Management, 77: 477-485. https://doi.org/10.1016/j.wasman.2018.04.033

[10] Ge, Z., Shi, H., Mei, X., Dai, Z., Li, D. (2016). Semi-automatic recognition of marine debris on beaches. Scientific Reports, 6(1): 25759. https://doi.org/10.1038/srep25759

[11] Miraliakbari, A., Sok, S., Ouma, Y. O., Hahn, M. (2016). Comparative evaluation of pavement crack detection using kernel-based techniques in asphalt road surfaces. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLI-B1, 2016 XXIII ISPRS Congress, Prague, Czech Republic, pp. 689-694. https://doi.org/10.5194/isprsarchives-XLI-B1-689-2016

[12] Zalama, E., Gómez-García-Bermejo, J., Medina, R., Llamas, J. (2014). Road crack detection using visual features extracted by Gabor filters. Computer ‐ Aided Civil and Infrastructure Engineering, 29(5): 342-358. https://doi.org/10.1111/mice.12042

[13] Chu, Y., Huang, C., Xie, X., Tan, B., Kamal, S., Xiong, X. (2018). Multilayer hybrid deep-learning method for waste classification and recycling. Computational Intelligence and Neuroscience, 2018: 5060857. https://doi.org/10.1155/2018/5060857

[14] Ministry of Construction of the People's Republic of China (1997). Urban Environmental Sanitation Quality Standard (1997, No.21).

[15] Mittal, G., Yagnik, K.B., Garg, M., Krishnan, N.C. (2016). SpotGarbage: smartphone app to detect garbage using deep learning. UbiComp '16: The 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, pp. 940-945. https://doi.org/10.1145/2971648.2971731

[16] Wei, S., Cheng, Z. (2017). Image-based garbage detection in urban scenes. Journal of Integration Technology. 39-52.

[17] Krizhevsky, A., Sutskever, I., Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25(2): 84-90. https://doi.org/10.1155/2018/5060857

[18] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L. (2018). Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[19] Ronneberger, O., Fischer, P., Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-

Assisted Intervention - MICCAI 2015, Munich, Germany, pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[20] Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z. (2014). Deeply-supervised nets. In Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA, pp. 562-570. https://doi.org/10.48550/arXiv.1409.5185

[21] Tomanek, D.P., Schrder, J. (2018). Value added heat map - Flächennutzung. Value Added Heat Map, pp. 31-44. https://doi.org/10.1007/978-3-658-16895-7_4