Hybrid-Supervised-Learning-Based Automatic Image Segmentation for Water Leakage in Subway Tunnels

Dongwei Qiu 1, Haorong Liang 1, Zhilin Wang 1, Yuci Tong 1 and Shanshan Wan 2,*

1 School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
2 School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
* Correspondence: wss@bucea.edu.cn

Abstract: Quickly and accurately identifying water leakage is one of the important components of the health monitoring of subway tunnels. A mobile vision measurement system consisting of several high-resolution, industrial, charge-coupled device (CCD) cameras is placed on trains to implement structural health monitoring in tunnels. Through the image processing technology proposed in this paper, water leakage areas in subway tunnels can be found and repaired in real time. A lightweight automatic segmentation approach to water leakage using hybrid-supervised-deep-learning technology is proposed. This approach consists of the weakly supervised learning Water Leakage-CAM and fully supervised learning WRDeepLabV3+. The Water Leakage-CAM is used for the automatic labeling of data. The WRDeepLabV3+ is used for the accurate identification of water leakage areas in subway tunnels. Compared with other end-to-end semantic segmentation networks, the hybrid-supervised learning approach can more completely segment the water leakage region when dealing with water leakage in complex environments. The hybrid-supervised-deep-learning approach proposed in this paper achieves the highest MIoU of 82.8% on the experimental dataset, which is 6.4% higher than the second. The efficiency is also 25% higher than the second and significantly outperforms other end-to-end deep learning approaches.

Keywords: mobile measurement; vision measurement; subway tunnel; structural health monitoring; deep learning; water leakage

1. Introduction

With the construction and operation of urban subways, various problems related to subway tunnels continue to emerge [1,2]. These have created serious social and economic problems. The main problems associated with subway tunnel diseases are shown in Table 1, in which water leakage is shown to be one of the main problems in subway tunnel diseases. The New York subway tunnels in the United States caused train delays due to water leakage caused by signal failure [3]. Among the 12 subway lines of Beijing in China, there are seven common subway tunnel diseases, including segment water leakage, cracks, misaligned lining, empty track bed, concrete deterioration and section ovalization. Water leakage is one of the main diseases in Beijing subway tunnels [4].

In order to ensure the safe operation of subways, the mobile vision-measurement system is often placed on trains for mobile vision measurement [5]. As shown in Figure 1, 10 high-resolution industrial charge-coupled device (CCD) cameras are placed on the top, left, and right sides and bottom of the train. With 10 LED lighting apparatuses, the subway tunnel was photographed in a ring, and the image data of the entire tunnel was obtained [6,7].

Table 1. Main diseases of subway tunnels.
The real-time mileage information of the train is obtained through the odometer, inertial navigation system, communication system, etc., and then the image with position information is obtained [8,9]. Additionally, real-time detection is carried out through a series of image-processing techniques [10,11]. When the water leakage is found, the relevant staff can repair it in time through the location information of the image.

In recent years, the development of deep learning has revolutionized visual measurement. For water leakage detection, semantic segmentation is a better choice. Object detection is divided into non-automatic detection and automatic detection. In non-automatic detection, the approach of fully supervised learning can more accurately segment the water leakage area [12,13]. Xue and Li [14] adopted full convolution (FCN) and Dai et al. [15] local-based FCN (R-FCN) to label, classify, and detect cracks and seepage. Huang et al. [16] adopted FCN Semantic segmentation of leaking water. However, due to its spatial invariance characteristics, small water leakage areas may be ignored [17]. Han et al.
[18] proposed a multi-spectral water-leakage detection method that integrates visual optics and thermal infrared and detects the water leakage through single-mode feature extraction and multi-mode feature fusion of FPN, but this method is only performed in laboratory simulations and is not suitable for complex environments. Xiong et al. [19] designed a deep learning system based on an image recognition model to detect water leakage. However, using a VIS color camera is greatly affected by the lighting conditions in the tunnel, and this reduces the detection accuracy. The fully supervised learning approach requires a large number of high-precision dataset labels, so the pre-label production time is long and the efficiency is low. The approach of weakly supervised learning is able to achieve semantic segmentation of objects with few data labels [20]. Zhu et al. [21] used weakly supervised networks to detect and segment different types of cracks and compared the segmentation effects of different networks but did not evaluate the accuracy, precision, and other indicators. Chang et al. [22] proposed a defect segmentation system based on weakly supervised learning, which improves the accuracy of classification and segmentation, but the efficiency is not high, and the problem of time-consuming data labeling has not been solved. Zhao et al. [23] used Mask R-CNN for instance segmentation of water leakage in tunnel linings. The time-consuming problem was solved without overfitting, but there was a problem of low segmentation accuracy. Wang et al. [24] proposed a new network framework as the backbone of the weakly supervised network and added the K-means clustering algorithm to improve the segmentation accuracy. However, due to the limitations of K-means, the image segmentation effect is not ideal in complex environments.

Automatic detection has become the mainstream trend of current research, which can ensure detection accuracy while ensuring efficiency. Zhang et al. [25] proposed a new method for automated pixel-level crack detection on 3D asphalt pavement and called it CrackNet-R. This method can achieve high precision in crack detection. However, it can easily create technical isolation for users in practical applications. Li et al. [26], Dung et al. [27], and Yang et al. [28] applied a fully convolutional network (FCN) to perform automatic pixel-level crack detection. This method has a good effect on large-pixel crack detection but has certain limitations for small cracks, and the effect is not ideal. Bang et al. [29] proposed an encoder–decoder network-based method for the automatic detection of road cracks at the pixel level. Transfer learning is performed using ResNet152, which achieves the best performance among several convolutional neural networks, but the experimental results do not achieve the expected accuracy. Liu et al. [30] propose the use of U-Net for automatic crack detection. U-Net has better robustness than Cha’s CNN method and is more suitable for crack detection. However, due to the complex environment and interference, it will generate redundant identification. Song et al. [13] proposed an improved automatic crack detection framework based on DeepLabV3+ for detection and obtained a speed of 23 frames per second, but the robustness is not strong, only for small data samples. Therefore, automatic object detection is still a challenge for pixel segmentation.

At present, there are still problems such as complex algorithms, long time consumption, and low detection accuracy in automatic detection. This paper proposes a lightweight hybrid supervised learning (HSL) approach that combines efficient weakly supervised learning with high-precision, fully supervised learning. Among them, weakly supervised learning is used for the automatic generation of water leakage pixel-level labels, while fully supervised learning is used for water leakage semantic segmentation. In weakly supervised learning, the Water Leakage-CAM (WL-CAM) method framework is proposed. Through a multi-level fusion strategy, high-precision pixel-level labels are obtained. First, an adaptive pixel segmentation clustering (APS) algorithm is used to generate image-level labels with less background information. Second, the WRes2Net deep learning network is used for training, and the class activation map (CAM) generated by training is used to generate label files and image-level labels for pixel value overlap and fusion, thereby reducing data labeling time and labor costs. In fully supervised learning,
this paper improves the DeepLabV3+ network, adding multiple learning rate decay methods, optimizers, and attention mechanisms. It can accurately segment water leakage in various complex environments, and the segmentation accuracy is greatly improved.

2. Approach

The HSL combines the advantages of weakly supervised learning and fully supervised learning. The accuracy of data labels is guaranteed without manual labeling, while the accuracy of fully supervised learning semantic segmentation is guaranteed.

2.1. Basic Idea

Figure 2 shows the basic idea of HSL. As a new approach for identifying water leakage in subway tunnels, it can automatically identify water leakage areas without labels.

Figure 2. Basic idea of HSL.

We take an unlabeled original dataset as an example to illustrate the implementation process of the HSL approach. The original dataset includes RGB images ($I_{RGB}$) and grayscale images ($I_{Gray}$). APS generates image-level labels ($I^{APS}$) in two steps. Firstly, the number $K$ of cluster types is calculated through the gray gradient distribution and gray peak value of $I_{Gray}$. Secondly, batch clustering is performed using the $I_{RGB}$ dataset to obtain $I^{APS}$. The WRes2Net deep learning network is used to extract the features of $I^{APS}$, generate a CAM feature map, and refine the CAM feature map to generate a CAM label ($I^{MD}_{S}$). The $I^{APS}$ and $I^{MD}_{S}$ are fused and superimposed. Based on the $I^{APS}$ category, the region with the largest category weight is reserved, and the pixel-level label $I$ is obtained. Through WRDeepLabV3+, the feature of water leakage is extracted for $I$, and the water leakage area is segmented.

2.2. Framework

As shown in Figure 3, the HSL approach adopts a two-stage structure, which is divided into a weakly supervised learning WL-CAM and a fully supervised learning WRDeepLabV3+. There are two steps in WL-CAM. Step 1, according to the characteristics of water leakage, is to propose the APS algorithm and the segment to obtain image-level labels. Step2 is to train WRes2Net semantic segmentation network to obtain a CAM [31]. The generated CAM will obtain the label file (Mask Data) through the random-walk algorithm (RW). Image-level tags are combined with it. To further improve the completeness and accuracy of dataset labels, pixel-level labels are obtained. Feature extraction and semantic segmentation of water leakage are carried out on pixel-level labels using WRDeepLabV3+. In WRDeepLabV3+, the main network is WRes2Net, ASSP is a spatial pyramid pooling module with atrous convolution, and transfer learning is performed through pixel-level labels and the original dataset. We use the channel attention mechanism (CA) to extract the high-level semantic information of water leakage, perform average pooling on the high-level semantic information to obtain a feature vector in the direction of the channel. Then use two nonlinear fully connected layers (FC) to obtain the cor-
relation relationship between channels to limit the complexity of the model. Finally Sigmoid is used for channel feature vector normalization. The spatial attention mechanism (SA) is used to focus the water leakage target area to extract the low-level semantic information of the water leakage, the semantic segmentation accuracy is improved through the global convolutional network, and finally, the channel feature vector normalization processing is performed using Sigmoid. Finally, through the combination of high-level semantic information and low-level semantic information, the water leakage area is generated.

**Figure 3.** The overall framework of the HSL.

2.3. Automatic Generation of Water Leakage Labels for Subway Tunnels

The automatic generation of water leakage labels for subway tunnels adopts the WL-CAM based on weakly supervised learning.

Image-level labels are generated through ASP, which is divided into the following steps:

1. Calculate the gray value of the image to find $K$. Suppose an image $I_{\text{gray}}$ consists of $N$ water basin characteristic areas, $I_{\text{gray}} = \{y_n|y_n \in \mathbb{R}^d, n = 1,2,3,...,N\}$, where $y_n$ consists of $M$ data that provide its characteristics. $y_n = \{\eta_m|\eta_m \in y_n, m = 1,2,3,...,M\}$, and $\eta_m$ is the gray value of $M$ pixels in the feature area. Calculate the grayscale histogram. In the grayscale histogram, $U_{\nu}^{\text{max}} = \{U_{\nu}^{\text{max}}|U_{\nu}^{\text{max}} \in I_{\text{gray}}, \nu = 1,2,3,...,V\}$, $U_{\nu}^{\text{max}}$ is the peak value of the gray value, and $\nu$ is the peak value number.

2. To further determine the number of clusters $K$, find the $y_n$ mean $\overline{y_n}$, and the number of types of $K$ is:
where \( a \) and \( b \) are grayscale values.

3. To process the RGB of the image, let \( I_{rgb} \) consist of \( N \) data, \( I_{rgb} = \{ S_n \}_{n=1}^{N} \). A binary variable \( P_{nk} \in \{0,1\} \) is introduced to indicate which \( k \) cluster \( S_k \) \((k = 1, 2, ..., K)\) any point \( S_n \) belongs to. When \( P_{nk} = 1 \), it means that the data belong to class \( k \); otherwise it is 0. The loss function is thus defined as:

\[
\Psi = \sum_{n=1}^{N} \sum_{k=1}^{K} P_{nk} \| S_n - S_k \|^2 \quad (3)
\]

\[
\| S_n - S_k \| = \sqrt{D_1^2 + D_2^2} \quad (4)
\]

\[
D_1 = \sqrt{(R_n - R_k)^2 + (G_n - G_k)^2 + (B_n - B_k)^2} \quad (5)
\]

\[
D_2 = \sqrt{(X_n - X_k)^2 + (Y_n - Y_k)^2} \quad (6)
\]

\( D_1 \) is the color distance, \( D_2 \) is the Euclidean space distance, \( \| S_n - S_k \| \) is the absolute value of the clustering distance. The APS converges the cluster centers through continuous iteration, and \( S_n \) and \( S_k \) can be optimized alternately during the iterative process. When \( K = 2 \), calculate the pixel area and keep the largest pixel area.

4. During clustering, the calculation of spatial distance will affect the classification of pixel points, resulting in the existence of independent pixels within the cluster. The Canny edge extraction algorithm obtains the closed cluster edge to determine the closed cluster. Calculate the color distance between the independent pixels inside the closed cluster and the \( K \) cluster centers, and assign it to the cluster area with the smallest color distance, as shown in Figure 4.

![Figure 4](image-url)
stability, the activation function ReLU is replaced by the Mish function [34]. The non-negative effect of the activation function can make the weight layer update less ideal, placing the activation layer before the normalization layer [33].

\[
I_{SM}^D = \sum_C a_S^C F_C
\]  

(7)

where \(I_{SM}^D\) is the Mask Data image, \(S\) is the label type, \(C\) is the number of channels, \(a_S^C\) is the weight of the \(C\)-th channel type \(t\), and \(F_C\) is the feature map of the \(C\)-th channel. Expand the weight \(a_S^C\) to obtain:

\[
a_S^C = \frac{1}{L} \sum_i \sum_j \frac{\partial \xi^S}{\partial F_C^{ij}}
\]  

(8)

where \(L\) is the width and height of the feature map, \(\xi^S\) is the probability of outputting the target type, and \(F_C^{ij}\) represents the pixel value at \((i, j)\) in the \(C\)-th feature map.

Since the weakly supervised semantic segmentation network only detects the main features of different categories in the image, it is difficult to obtain a complete object response map. Combining Mask Data with image-level labels:

\[
I = aI^{APS} + \beta I_{SM}^D + \eta
\]  

(9)

\(I\) is the pixel-level label, \(I^{APS}\) is the image-level label, \(a\) and \(\beta\) are the weights, and \(\eta\) is the residual.

Based on \(I^{APS}\), \(I^{APS}\) contains \(n\) subclass labels, which overlap with \(I_{SM}^D\), and the subclass label with the highest overlap weight value is the feature target. The images are corrected and supplemented by the residual network in the training network.

2.4. Semantic Segmentation of Water Leakage in Subway Tunnels
As can be seen from Figure 6, WRDeepLabV3+ consists of two modules. The encoding area (Encoder) is used for the extraction of high-level semantic information. The decoding area (Decoder) is used for the extraction of low-level semantic information. The WRes2Net network proposed in 2.3 is used as the backbone network of WRDeeplabv3+ to extract the feature information of water leakage. The channel attention mechanism (CA) is added to the Encoder [35]. CA assigns larger weights to highly responsive channels after feature extraction using depth-wise-separable convolutional layers of different channels. Suppose the high-level semantic information is \( F^h_c \in I^{W \times H \times C} \), \( F^h_c = \{ F_1, F_2, ..., F_C \} \) where \( W \) and \( H \) are the width and height of the input feature image, and \( C \) is the channel number.

CA is represented as:

\[
    f_{CA}(g_c, \Phi_{CA}) = e_1\{fc_2[r(fc_1(g_c, \Phi_{CA1}), \Phi_{CA2})]\}
\]

(10)

In the formula, \( g_c \) is the feature map after \( F^h_c \) average pooling, \( \Phi_{CA} \) is the parameter in the channel attention module, \( e_1 \) is the Sigmoid activation, \( fc \) is the fully connected layer, and \( r \) is the ReLU activation. The CA module outputs \( f_{CA} \) and weights the feature map to obtain the output feature map:

\[
    F^{hr}_c = f^h_c \cdot f_{CA}
\]

(11)

Compared with the encoding area, the decoding area can obtain the location and edge information of the target information and other features. However, there is a lot of background information, which will affect the segmentation accuracy to a certain extent. The spatial attention mechanism (SA) is introduced in the Decoder. It can focus on the target feature area, adaptively combine high-level features with low-level features, and use high-level features to filter out background information [36]. To obtain global information without increasing parameters, semantic segmentation is improved by global convolutional networks. Two-layer convolution operation is used, and the two-layer convolution kernels are \( 1 \times 5 \) and \( 5 \times 1 \), respectively, which are used to obtain key feature information. SA is expressed as:

\[
    A_1 = Con_1[Con_2(F^h_c, \Phi_{SA1}), \Phi_{SA2}]
\]

(12)

\[
    A_2 = Con_2[Con_1(F^h_c, \Phi_{SA1}), \Phi_{SA2}]
\]

(13)

\[
    f_{SA}(F^{hr}_c, \Phi_{CA}) = e_2(A_1 + A_2)
\]

(14)

In the formula, \( A_1 \) is the feature map completed by convolution of convolution kernels \( 1 \times 5 \) and \( 5 \times 1 \), \( A_2 \) is the feature map completed by convolution of convolution kernels \( 5 \times 1 \) and \( 1 \times 5 \), \( Con_1 \) is the convolution kernel of \( 5 \times 1 \times C \), \( Con_2 \) is the convolution kernel of \( 1 \times 5 \times C \), \( e_2 \) is the Sigmoid activation, \( F^{hr}_c \) is obtained by SA weighting, and \( \Phi_{CA} \) is the parameter in the spatial attention module.

Training a complex deep learning model can take a long time, and the optimizer can improve the training efficiency of the model. At the same time, different optimizers can also improve the performance of the model and achieve better training results. In the training process, the learning-rate decay methods StepLR and CosineAnnealingLR (CosLR) are added, and the two optimizers SGD and Adam are also compared.

In addition, the training is divided into a freezing phase (Freeze) and an unfreezing phase (Unfreeze). At the same time, the Focal Loss function is used to solve the problem of positive and negative sample imbalance [37]. Its formula is shown in (15).

\[
    \text{LOSS} = \begin{cases} 
        -\kappa(1-y)^\delta, & y = 1 \\
        -(1-\kappa)y^\delta \log(1 - y'), & y = 0 
    \end{cases}
\]

(15)

where \( y' \) denotes the output after the activation function, \( \kappa \) denotes the loss weight of the balanced factor-regulated samples, and the sum of the loss weights of all categories is 1. \( \delta \) denotes the balanced factor-controlled hard and easy sample loss, and \( \delta \geq 0 \). When \( \delta = 0 \), the focal loss function degenerates into the ordinary cross-entropy loss function with \( \kappa \). When \( \delta \) increases, the model pays more attention to the hard-to-distinguish samples.
3. Experiment

In this study, the hardware of the experiment is a desktop computer, a mobile visual measurement system. The software is Python 3.7, Pytorch 1.10.2 and Labelme. Computer configuration as CPU: AMD Ryzen 7 5800X with Radeon Graphics 4.6 GHz. GPU: NVIDIA GeForce RTX 3060 6 GB. RAM: 16 GB. The mobile vision measurement system is placed on the left and right sides and the top of the subway locomotive. The train takes a circular photograph of the subway tunnels to obtain water leakage data. The mobile vision-measurement system includes CCD industrial cameras, LED lighting equipment, power supplies, the odometer, the inertial navigation system, the communication system, etc. The CCD industrial camera parameters are as follows: pixels, 3.5 million; pixel size, 3.75 μm × 3.75 μm; target size, 1/3″; and frame rate: 400–5250 fps.

3.1. Dataset

The experimental dataset consists of two parts, as shown in Table 2. The first part (a) is the water leakage dataset of Shanghai subway tunnels in China [38]. The second part (b) is the dataset of Beijing subway tunnels in China collected by our mobile vision-measurement system.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Collection Equipment</th>
<th>Number of Images</th>
<th>Resolutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>MTI-200a</td>
<td>3555</td>
<td>1600 × 1200 and 1944 × 2592</td>
</tr>
<tr>
<td>b</td>
<td>Mobile vision measurement system</td>
<td>6739</td>
<td>2560 × 1440</td>
</tr>
</tbody>
</table>

After data screening, the experimental dataset consists of 6000 images, including the subway tunnel environment, different types of water leakage, and other diseases. The size of each image is 512 × 512. The dataset is divided into the Train dataset and the Test dataset at a ratio of 4:1.

In order to compare the segmentation effect of automatic synthetic labels and manual labels, the data are divided into two groups (as shown in Table 3). The group A is automatic synthetic labels (AL-Mask images). The group B is artificial labels (Ground truth). Ground truth is labeled with Labelme and is saved in the JSON format.
Table 3. Classification of the experimental dataset.

<table>
<thead>
<tr>
<th>Groups</th>
<th>Train</th>
<th>Val</th>
<th>Test</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4000</td>
<td>1000</td>
<td>1000</td>
<td>6000</td>
</tr>
<tr>
<td>B</td>
<td>4000</td>
<td>1000</td>
<td>1000</td>
<td>6000</td>
</tr>
</tbody>
</table>

3.2. Experimental Scheme

Firstly, we demonstrate the excellent segmentation ability of APS in weakly supervised learning and the method’s effectiveness in automatically generating water leakage labels. First, the original image dataset is segmented into image-level labels via APS. As the subway tunnels mostly have a gray and white background, it forms a strong color contrast with the water leakage. The images are converted to grayscale images, the water leakage area is dark in the grayscale images, and the grayscale value is low. Through a large number of analyses and statistics of the gray value distribution of the water leakage area, it is found that the gray value of the water leakage area is greater than 60 and less than 130, while the background area is greater than 130 [39,40]. Through global threshold segmentation, the background information with a gray value greater than 130 is filtered out and then segmented and compared with images of other image segmentation algorithms. So a is chosen as 60, and b is chosen as 130. Second, image-level labels for multi-class training are provided via WRNet. CAM label files are generated for fusion with image-level labels. Redundant background information is removed, and AL-Mask images are generated and compared with the ground truth. After many experiments, the fusion weights α and β are taken as 0.8 and 0.2, respectively.

Secondly, the robustness of fully supervised learning WRDeepLabV3+ is demonstrated. Deep learning training on AL-Mask images is carried out using WRDeepLabV3+. Due to the complex environment of subway tunnels and different noise information from water leakage, the water leakage data are divided into five categories, and the semantic segmentation effect of WRDeepLabV3+ is displayed through different categories.

Finally, the advanced nature and accuracy of the HSL approach are demonstrated. The performance of the approach combined with WL-CAM and WRDeepLabV3+ is compared with other advanced semantic segmentation networks, including SC-CAM [22], UNet [41], PSPNet [42], HRNet [43], DeepLabV3+ [44], and WRDeepLabV3+. Because of the complex environment of ground tunnels, water leakage has the following three characteristics [45]. First, due to the lateral extrusion and gravity of the lining, as well as the influence of different lining gaps, the water leakage will occur on the lining surface, forming strip water leakage (including horizontal and vertical directions) and blocking water leakage. Second, there will be oil stains and artificial marks on the surface of the inner lining, which will be similar to the color and shape of the water leakage, affecting the segmentation accuracy. Third, the water leakage area will be covered by meter boxes, cables, pipes, etc., causing interference. The water leakage data are divided into five categories, using the HSL approach proposed in this paper and other end-to-end semantic segmentation networks including EM [46], CRF-RNN [47], 1Stage [48], and AA&LR [49] for semantic segmentation comparison is also performed on the dataset.

The experiments are comprehensively evaluated by using Precision, Recall, IoU, MIoU, and F1 to judge the performance of the model by researchers [50].

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{16}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{17}
\]

\[
\text{IoU} = \frac{TP}{TP + FP + FN} \tag{18}
\]
\[ MIoU = \frac{TP}{TP + FP + FN} + \frac{TN}{TN + FN + FP}/2 \]  

\[ F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \]  

where TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative.

3.3. Experimental Results

The segmentation results of APS are compared with several other image-segmentation algorithms (as shown in Table 4). K-means and Otus have more background information than global threshold segmentation (threshold is 90). For the water leakage area, the first three segmentation algorithms all have the phenomenon of incomplete segmentation, and independent pixels appear. Compared with the three segmentation algorithms, APS can better segment the water leakage area. The background information is effectively removed while retaining the water leakage area.

<table>
<thead>
<tr>
<th>Effects</th>
<th>K-Means</th>
<th>Otus</th>
<th>Threshold Segmentation</th>
<th>APS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water leakage details</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Background information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A comparison of the labeling effect of ground truths and AL-Mask images is shown in Table 5. The first column shows the original dataset of water leakage. The second column shows the water leakage dataset (ground truth) manually marked by Labelme. The third column shows the image-level water leakage dataset obtained by APS. The fourth column shows the mask data, and the fifth column shows the final pixel-level water leakage dataset (AL-Mask images). Compared with ground truth, the edge of the AL-Mask images is not smooth, but it can effectively extract the water leakage area.
Table 5. Comparison of the manually labeled dataset and the automatically labeled dataset.

<table>
<thead>
<tr>
<th>Original Images</th>
<th>Ground Truths</th>
<th>APS Images</th>
<th>Mask Data</th>
<th>AL-Mask Images</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
</tbody>
</table>

The subway tunnel’s surface environment is complex, and there are various types of noise, such as the occlusion of objects and the influence of artificial or oil stains, which cause certain difficulties in the segmentation of water leakage. As shown in Table 6, five typical noise images of water leakage of subway tunnels are listed, as well as their corresponding ground truth, WL-CAM, and WRDeepLabV3+ renderings. Under the complex background, WL-CAM has over-segmentation and incomplete segmentation, while WRDeepLabV3+ obtains a better segmentation effect.

Table 6. Water leakage segmentation effects of WL-CAM and WRDeepLabV3+ on different noises.

<table>
<thead>
<tr>
<th>Noise Information</th>
<th>Iron Tubes</th>
<th>Cables</th>
<th>Iron Wires</th>
<th>Electricity Meters</th>
<th>Stains</th>
</tr>
</thead>
<tbody>
<tr>
<td>The original images</td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
</tr>
<tr>
<td>Ground truths</td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
</tr>
<tr>
<td>WL-CAM</td>
<td><img src="image21.png" alt="Image" /></td>
<td><img src="image22.png" alt="Image" /></td>
<td><img src="image23.png" alt="Image" /></td>
<td><img src="image24.png" alt="Image" /></td>
<td><img src="image25.png" alt="Image" /></td>
</tr>
</tbody>
</table>
As shown in Table 7, the HSL approach proposed in this paper is compared with the current advanced semantic segmentation network. Since WL-CAM adopts two-stage training, a larger learning rate and batch size are used in the freezing stage, and a smaller learning rate and batch size are used in the thawing stage, which greatly shortens the training time. Compared with SC-CAM, WL-CAM is more time-saving, and all indicators are better than SC-CAM. In fully supervised learning, PSPNet has the highest efficiency, but the accuracy is sacrificed. WRDeepLabV3+ based on WRes2Net101 optimizes the learning rate and adds an optimizer so that the first five evaluation indicators are ahead of other fully supervised networks. It can be seen that the HSL approach proposed in this paper can achieve comparable or even better performance than the existing automatic annotation method SC-CAM. Its segmentation accuracy can also be comparable to the full supervised network. Since the HSL approach includes multiple steps, other models only calculate the model training time, so the time used by this approach is not comparable.

As shown in Table 8, the segmentation effects of the HSL approach and the end-to-end semantic segmentation network under different water leakage types are compared. It can be seen from this that, compared with block water leakage, EM, 1Stage, and AA&LR have better segmentation results. They can identify small water leakage blocks. While the approach in this paper can more completely segment small water leakage blocks. For vertical water leakage, the EM segmentation is incomplete, and the CRF-RNN is over-fitting. For horizontal water leakage, the EM has the problem of excessive segmentation. For water leakage with stain interference, redundant background information appears in the EM and 1Stage segmentation regions. However, in the first four approaches, the outline of the segmentation area is lacking or overfitting. The approach in this paper segmented the region with a more refined and complete outline. For the occluded water leakage, when the occluded part cannot be known, the five approaches all show good results, but EM still has the problem of incomplete segmentation.

As can be seen from Table 9, the MIoU of the approach proposed in this paper outperforms other end-to-end semantic segmentation networks on both the Val dataset and the Test dataset, reaching 81.7% and 82.8%, respectively. HSL is also the most efficient, 25% higher than the second.

**Table 7.** Comparison of evaluation results of different semantic segmentation under the proposed dataset.

<table>
<thead>
<tr>
<th>Method Categories</th>
<th>Methods</th>
<th>Loss</th>
<th>MIoU (%)</th>
<th>Recall (%)</th>
<th>F1 (%)</th>
<th>Precision (%)</th>
<th>Epoch</th>
<th>Training Time</th>
<th>Testing Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weakly supervised learning</td>
<td>SC-CAM</td>
<td>0.154</td>
<td>69.1</td>
<td>63.5</td>
<td>70.2</td>
<td>78.7</td>
<td>50</td>
<td>18 h 58 m 07 s</td>
<td>0.087s/image</td>
</tr>
<tr>
<td></td>
<td>WL-CAM</td>
<td>0.132</td>
<td>73.6</td>
<td>75.6</td>
<td>79.5</td>
<td>83.9</td>
<td>50</td>
<td>15 h 21 m 16 s</td>
<td>0.061s/image</td>
</tr>
<tr>
<td></td>
<td>PSPNet</td>
<td>0.096</td>
<td>80.4</td>
<td>82.8</td>
<td>80.1</td>
<td>77.6</td>
<td>100</td>
<td>11 h 04 m 46 s</td>
<td>0.028s/image</td>
</tr>
<tr>
<td></td>
<td>U-Net</td>
<td>0.077</td>
<td>82.0</td>
<td>83.8</td>
<td>81.9</td>
<td>80.1</td>
<td>100</td>
<td>17 h 55 m 59 s</td>
<td>0.074s/image</td>
</tr>
<tr>
<td></td>
<td>HRNet</td>
<td>0.089</td>
<td>81.0</td>
<td>83.3</td>
<td>80.8</td>
<td>78.6</td>
<td>100</td>
<td>19 h 57 m 58 s</td>
<td>0.095s/image</td>
</tr>
<tr>
<td></td>
<td>DeepLabV3+</td>
<td>0.079</td>
<td>81.8</td>
<td>85.2</td>
<td>81.8</td>
<td>78.7</td>
<td>100</td>
<td>11 h 53 m 34 s</td>
<td>0.033s/image</td>
</tr>
<tr>
<td></td>
<td>WRDeepLabV3+</td>
<td>0.053</td>
<td>86.2</td>
<td>88.6</td>
<td>86.7</td>
<td>84.9</td>
<td>100</td>
<td>13 h 22 m 33 s</td>
<td>0.035s/image</td>
</tr>
<tr>
<td>HSL (proposed)</td>
<td></td>
<td>0.084</td>
<td>82.8</td>
<td>85.2</td>
<td>82.9</td>
<td>80.6</td>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8. Comparison of segmentation effects of HSL and the end-to-end semantic segmentation network under different water leakage types.

<table>
<thead>
<tr>
<th></th>
<th>Blocks</th>
<th>Vertical Strips</th>
<th>Horizontal Strips</th>
<th>Stains</th>
<th>Occlusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>The original images</td>
<td>EM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CRF-RNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA&amp;LR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HSL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9. MIoU and the efficiency comparison between the HSL and the end-to-end semantic segmentation network on the proposed dataset.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Backbone</th>
<th>Val (%)</th>
<th>Test (%)</th>
<th>Testing time</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>VGG16</td>
<td>58.4</td>
<td>59.8</td>
<td>0.092 s/image</td>
</tr>
<tr>
<td>CRF-RNN</td>
<td>VGG16</td>
<td>61.6</td>
<td>62.9</td>
<td>0.068 s/image</td>
</tr>
<tr>
<td>1Stage</td>
<td>WideResNet38</td>
<td>73.4</td>
<td>74.1</td>
<td>0.044 s/image</td>
</tr>
<tr>
<td>AA&amp;LR</td>
<td>WideResNet38</td>
<td>75.6</td>
<td>76.4</td>
<td>0.052 s/image</td>
</tr>
<tr>
<td>HSL</td>
<td>WRes2Net101</td>
<td>81.7</td>
<td>82.8</td>
<td>0.035 s/image</td>
</tr>
</tbody>
</table>

4. Analysis and Discussion

This section mainly discusses in detail the advanced nature of the data label automatic labeling method Water Leakage-CAM and the water leakage semantic segmentation network WRDeepLabV3+ included in the HSL approach.

4.1. Performance Evaluation of Water Leakage—CAM

In WL-CAM, the quality of image-level labels depends on APS, and the quality of the pixel-level label depends on the feature extraction network WRes2Net.

WRes2Net is crucial for pixel-level labeling, which is related to the quality of the dataset and the accuracy of the subsequent training of fully supervised segmentation models. This paper chooses to improve based on Res2Net101, which has a deeper feature ex-
traction convolution layer and stronger multi-scale convolution ability. Through the improvement of this model, WRes2Net101 is obtained. In order to compare the performance impact of the two segmentation models on WL-CAM, they are both trained on the same dataset and judged with the same criteria. The MIoU of WL-CAM based on WRes2Net101 is 6.3% higher than that of WL-CAM based on Res2Net101. The overall segmentation accuracy and performance are better than those of WL-CAM based on Res2Net101, as shown in Table 10 and Figure 7.

Table 10. Performance evaluation of WL-CAM with different backbone networks based on the proposed dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>MIoU (%)</th>
<th>F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WL-CAM (Res2Net101)</td>
<td>71.4</td>
<td>74.3</td>
<td>67.3</td>
<td>72.8</td>
</tr>
<tr>
<td>WL-CAM (WRes2Net101)</td>
<td>83.9</td>
<td>75.6</td>
<td>73.6</td>
<td>79.5</td>
</tr>
</tbody>
</table>

Figure 7. Loss of WL-CAM with different backbone networks.

4.2. Performance Evaluation of WRDeepLabV3+ Semantic Segmentation Network

WRDeepLabV3+ uses the WRes2Net101 structure as the core network with an Epoch of 50 in the freezing phase and 50 in the thawing phase. Firstly, the effect of the optimizer on WRDeepLabV3+ (without attention mechanism) under different learning rate decay methods is compared. As shown in Figure 8, the initial learning rate is 0.0005, the green line is the MIoU of Adam under CosLR, the blue line is the MIoU of Adam under StepLR, the yellow line is the MIoU of SGD under CosLR, and the orange line is the MIoU of SGD under StepLR. Under different learning-rate-decay methods, the MIoU of the optimizer under CosLR is higher than that of StepLR. Experiments show that the proposed model converges faster and can obtain the best segmentation effect when using Adam optimizer and CosLR decay. Figure 9 shows that the initial learning rate is 0.0005 using Adam optimizer and CosLR decay. Compared with the training model of WRDeepLabV3+ and DeepLabV3+, MIoU is improved by 3.3%.
Secondly, we compare WRDeepLabV3+ without and with the attention mechanism. As shown in Table 11, under the same dataset, due to the addition of channel attention mechanism and spatial attention mechanism, Loss is reduced by 7.5%, Recall is improved by 2.4%, Precision is improved by 3%, and IoU is improved by 2.2%.

<table>
<thead>
<tr>
<th>WRDeepLabV3+</th>
<th>Loss</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
<th>IoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No attention mechanism</td>
<td>0.057</td>
<td>86.2</td>
<td>81.9</td>
<td>75.4</td>
</tr>
<tr>
<td>Add attention mechanism</td>
<td><strong>0.053</strong></td>
<td><strong>88.6</strong></td>
<td><strong>84.9</strong></td>
<td><strong>77.6</strong></td>
</tr>
</tbody>
</table>

Finally, the effect of transfer learning on the performance of WRDeepLabV3+ is discussed. The transfer learning of the backbone network model is to use the parameters of the trained model (pre-trained model) to transfer to the new model to help the training of the new model. As shown in Figure 10, it allows the network to improve faster during training. After training, the convergence effect and performance of the model are better.
In this paper, an HSL approach for the automatic segmentation of water leakage images is proposed. The water-leakage labels are automatically generated by WL-CAM based on the weakly supervised method. The water leakage is semantically segmented based on the fully supervised method WRDeepLabV3+.

In WL-CAM, according to the characteristics of water leakage, an adaptive APS algorithm is proposed, which can accurately and completely segment the water leakage area and generate image-level labels. Secondly, a weakly supervised network WRes2Net is proposed to generate CAM labels. The CAM label files overlap with image-level labels and pixel values and are fused to generate pixel-level labels, which further improves label accuracy and saves manual labeling costs.

In WRDeepLabV3+, the WRes2Net in WL-CAM is used as the core network of WRDeepLabV3+. WRDeepLabV3+ based on a fully supervised network is proposed, and the parameters and framework are adjusted to improve the performance. The excellent robustness of WRDeepLabV3+ is verified by semantic segmentation of water leakage images with different complex noise information.

WL-CAM and WRDeepLabV3+ were tested and compared with other state-of-the-art semantic segmentation methods on the proposed dataset. The results show that WL-CAM has better performance of automatically generating labels than other methods, with a difference of up to 12.1% in the evaluation indicators. All performance indicators of WRDeepLabV3+ are also ahead of other advanced fully supervised methods.

The HSL automatic segmentation approach for water leakage images performs on par with WRDeepLabV3+ and the manual labeling dataset, validating the feasibility and accuracy of WL-CAM. Due to the excellent performance of WL-CAM and WRDeepLabV3+, the hybrid-supervision approach achieves an MIoU of 82.8% on the dataset, which is 6.4% higher than the second end-to-end approach. The efficiency is 25% higher than that of the second. In a complex subway tunnel environment, it can accurately segment the water leakage area.

In the future, more images of water leakage in subway tunnels will be collected as the training dataset to improve the segmentation accuracy and performance of the HSL approach. At the same time, the weakly supervised learning method will be optimized to reduce the complexity of the method and save the training time of the dataset.

Author Contributions: D.Q.: the conception and design of the work, data analysis, problem modeling. H.L.: data acquisition, problem modeling. Z.W.: data analysis, problem modeling. Y.T.: data acquisition, problem modeling. S.W.: the conception and design of the work, methodology, problem modeling, funding acquisition. All authors have read and agreed to the published version of the manuscript.
Funding: the National Natural Science Foundation of China (No.61902016), the postgraduate educational and teaching quality improvement project of BUCEA, China (No.J2022005), the BUCEA Post Graduate Innovation Project (No.PG2022118).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

References


