Intelligent Detection of Tunnel Leakage Based on Improved Mask R-CNN

Wenkai Wang¹, Xiangyang Xu¹,* and Hao Yang²

¹ School of Rail Transportation, Soochow University, Suzhou 215006, China
² School of Transportation and Civil Engineering, Nantong University, Nantong 226019, China
* Correspondence: x.y.xu@suda.edu.cn

Abstract: The instance segmentation model based on deep learning has addressed the challenges in intelligently detecting water leakage in shield tunneling. Due to the limited generalization ability of the baseline model, occurrences of missed detections, false detections, and repeated detections are encountered during the actual detection of tunnel water leakage. This paper adopts Mask R-CNN as the baseline model and introduces a mask cascade strategy to enhance the quality of positive samples. Additionally, the backbone network in the model is replaced with RegNetX to enlarge the model’s receptive field, and MDConv is introduced to enhance the model’s feature extraction capability in the edge receptive field region. Building upon these improvements, the proposed model is named Cascade-MRegNetX. The backbone network MRegNetX features a symmetrical block structure, which, when combined with deformable convolutions, greatly assists in extracting edge features from corresponding regions. During the dataset preprocessing stage, we augment the dataset through image rotation and classification, thereby improving both the quality and quantity of samples. Finally, by leveraging pre-trained models through transfer learning, we enhance the robustness of the target model. This model can effectively extract features from water leakage areas of different scales or deformations. Through instance segmentation experiments conducted on a dataset comprising 766 images of tunnel water leakage, the experimental results demonstrate that the improved model achieves higher precision in tunnel water leakage mask detection. Through these enhancements, the detection effectiveness, feature extraction capability, and generalization ability of the baseline model are improved. The improved Cascade-MRegNetX model achieves respective improvements of 7.7%, 2.8%, and 10.4% in terms of AP, AP₀.₅, and AP₀.₇₅ compared to the existing Cascade Mask R-CNN model.

Keywords: tunnel water leakage; Mask R-CNN; Cascade; RegNetX; MDConv

1. Introduction

With the popularization of urban subways, maintaining the safe and long-lasting operation of subways has become the focus of social attention. The safe operation of the subway has a close relationship with the tunnel, as it plays an important role in the subway system. However, due to the complexity of the internal environment and geological conditions of tunnels, various defects often appear inside tunnels, which seriously affect tunnel safety [1]. Among these tunnel defects, tunnel water seepage is one of the most common defects, and the accumulation of tunnel water seepage reduces the load-bearing capacity of the tunnel [2]. In urban life, the subway, as one of the means of transportation for residents to commute to work, can be paralyzed by water leakage problems that cause the subway to stop running. As one of the transportation modes for urban residents to commute, the suspension of subway operation due to tunnel water leakage can paralyze urban transportation. For example, commuters using Line 1 in St. Petersburg, Russia experienced inconvenience due to the suspension of operation caused by the accumulation of tunnel water leakage [3]. Serious tunnel water leakage can lead to tunnel safety accidents.
For example, the Foshan subway in Guangdong Province experienced multiple casualties due to tunnel water leakage issues [4]. Therefore, to detect tunnel leakage in time and maintain the safety inside the tunnel, it is crucial to know how to detect subway tunnel leakage efficiently and intelligently [5]. The traditional method of water leakage detection usually relies on manual visual inspection, which is not only inefficient but also requires a significant amount of labor [6]. In recent years, with the development of laser scanning technology and high-speed cameras, mainstream internal tunnel inspection methods have increasingly employed laser scanning technology to build a point cloud model of the interior of the tunnel, as well as high-speed cameras to capture 2D tunnel leakage images [7–9]. By observing the point cloud model inside the tunnel, the overall safety condition of the tunnel can be viewed soon enough for timely maintenance and protection. Through the acquired 2D tunnel images, intelligent detection of tunnel water leakage areas can be carried out. The application of deep learning in computer vision for image detection performs well [10,11].

The main progress of automated image detection is attributed to the development of convolutional neural networks (CNN). CNNs have demonstrated excellent performance in many fields such as image object detection and image instance segmentation [12,13]. With the continuous advancement of convolutional neural networks, the model known as Regions with CNN (R-CNN) has been introduced as a two-step object detection framework. It utilizes the selective search method to generate multiple candidate object regions and performs object detection based on this. It was first introduced for object detection. He et al. introduced SPP-NET by incorporating spatial pyramid pooling (SPP) into R-CNN, which not only reduces the impact of input image size on the network but also improves the accuracy of object detection [14]. Girshick further integrated region-of-interest pooling (RoI Pool) into SPP-NET, proposing an improved algorithm called Fast R-CNN, which significantly enhances the speed of object detection [15]. Ren et al. replaced the slow Selective Search algorithm used for candidate object region generation with a new region proposal network (RPN), leading to the development of Faster R-CNN [16]. Currently, Faster R-CNN has become the mainstream algorithm in the field of object detection due to its superior performance [17]. To accurately extract object features from detection images and segment detected objects, Long et al. proposed the fully convolutional network (FCN), which is the first end-to-end fully convolutional network for pixel-level prediction [18]. It has been applied in various fields such as tunnel and road disease detection, achieving significant results [19]. Mask R-CNN is proposed for instance segmentation, which combines the Faster R-CNN framework and FCN algorithm to perform both object detection and semantic segmentation tasks [20]. The introduction of Mask R-CNN has led to the application of instance segmentation across various domains, combining the Faster R-CNN framework and FCN algorithm to perform both object detection and semantic segmentation tasks simultaneously. Deep-learning-related fault detection has been widely used in various industries, proving the feasibility of deep learning tunnel leakage detection [21,22]. Mask R-CNN is regarded as one of the state-of-the-art models for solving object detection and instance segmentation tasks, capable of learning rich features from input images. Zhao et al. utilized Mask R-CNN for instance segmentation of water leakage in shield tunnels [23]. Xue et al. made improvements to Mask R-CNN for water leakage detection in tunnels, enhancing both the accuracy and speed of detection [24]. However, the direct use of the baseline model Mask R-CNN for detecting tunnel water leakage will result in detection errors. In this paper, by improving the baseline model, we propose an instance segmentation detection model that is more suitable for irregular regions like tunnel water leakage.

In the actual detection process, directly applying the baseline model to tunnel water leakage detection may suffer from poor generalization, leading to issues such as missed detections, false positives, and duplicate detections. Moreover, for irregular regions like water leakage in tunnels, the baseline model may have a limited receptive field, resulting in inadequate feature extraction. This study primarily focuses on improving the generalization of the model by enhancing the baseline model to achieve better tunnel water
leakage detection. Traditional convolutional methods may encounter challenges in extracting features from irregular areas like tunnel water leakage, including limited receptive fields and insufficient capability to extract edge features of tunnel leakage. This study introduces new backbone networks and convolutional methods to enhance the model’s feature extraction capabilities. By improving the Mask R-CNN model, this study proposes a tunnel water leakage detection model with a larger receptive field and stronger feature extraction capabilities, thereby achieving high-precision mask detection of tunnel water leakage. Section 2 focuses on the baseline model Mask R-CNN, cascade strategy, RegNetX, deformable convolution, and the new model Cascade-MRegNetX. Section 3 describes the preprocessing measures for the dataset of this paper, including dataset expansion and classification. Section 4 describes the experimental configuration, experimental evaluation metrics, and model training. Section 5 analyzes the comparison of the model proposed in this paper with other backbone network models and the improvement of the model by adding migration learning. Section 5.3 presents the experimental results, which are divided into two main areas: (1) error analysis and (2) accuracy analysis. Section 6 summarizes the whole paper.

2. Network Structure

2.1. Mask R-CNN

A two-stage instance segmentation model based on deep learning has been developed to efficiently detect water leakage in shield tunnels, addressing the challenge of manual detection. The classical Mask R-CNN model, commonly used for two-stage instance segmentation, serves as the baseline model. He et al. introduced a small fully convolutional network (FCN) to add mask branches to the Faster R-CNN structure, achieving instance segmentation prediction. This network framework is illustrated Figure 1 [20]. Zhao proposed Cascade Mask R-CNN, which consists of a series of detectors trained with IoU thresholding [25]. This approach involves cascading the refinement of tasks individually, with each detector predicting independently, and then integrating the prediction results. Due to the limited number and quality of positive samples obtained by the baseline model, this paper incorporates a cascade strategy to enhance the baseline model’s ability to acquire positive samples.

![Mask R-CNN network framework.](image)

As shown in Figure 1, Mask R-CNN mainly consists of three parts: the backbone, the neck, and the head. The structure of the baseline model in this paper is composed of the following three parts.

(1) Backbone: ResNet-50 + FPN (Feature Pyramid Networks)

ResNet, being a classical backbone network for feature extraction, generates five layers of feature maps with sizes of \( \left( \frac{H}{2}, \frac{W}{2} \right) \), \( \left( \frac{H}{4}, \frac{W}{4} \right) \), \( \left( \frac{H}{8}, \frac{W}{8} \right) \), \( \left( \frac{H}{16}, \frac{W}{16} \right) \), and \( \left( \frac{H}{32}, \frac{W}{32} \right) \), respectively. Each layer focuses on different types of information extraction, where low-level features capture detailed information, while high-level features extract semantic information. However, the high-level features often suffer from a loss of spatial resolution.
To address this issue, the feature pyramid network (FPN) structure is incorporated into the backbone network of Mask R-CNN to fuse high- and low-layer features. In FPN, high-level features with low resolution are dimensionally reduced through convolution, then up-sampled to match the size of the previous feature map, and finally fused through element-wise summation. RPN is the region-generating network, which is used for the extraction of the candidate frames.

(2) Neck: RoI Align

In the Mask R-CNN model, RoI Align is employed instead of RoI pooling due to the large quantization error associated with RoI pooling. RoI Align utilizes bilinear interpolation to standardize the sizes of different input RoIs, resulting in an output feature map of uniform size. The process involves dividing the input RoI into $2 \times 2$ cells (shown in red in Figure 2), with each cell serving as a sampling point. The coordinates of these sampling points are represented as floating-point numbers, and bilinear interpolation (shown by pink arrows in Figure 2) is applied to obtain the values of the target pixel points. Finally, max pooling is performed on the four sampling points to obtain the final result.

![Figure 2. RoI Align.](image)

(3) Head: FCN (Fully Convolutional Networks) + FC (Fully Connected).

The header network primarily comprises bounding box recognition (classification and regression) and mask prediction. The fully connected (FC) layer is responsible for classification through feature extraction, where the fully convolutional network (FCN) is employed for semantic segmentation of the image and end-to-end pixel-level prediction. The final output includes the C1, B1, and S1 parts, as shown in Figure 1.

Mask R-CNN is a model that extends Faster R-CNN by adding mask branches on top of Faster R-CNN. While Faster R-CNN handles target detection, the Mask branches handle semantic segmentation. The process begins with the input image undergoing feature extraction by the backbone network ResNet-50 and feature pyramid network (FPN), producing different feature layers. These features are then used to generate regions of interest (RoIs) via the region proposal network (RPN). The RoIs selected by RPN are aligned with the backbone network’s feature map using RoI Align. Finally, the aligned RoIs are fed into the head network to predict the bounding box location and mask segmentation of the target region. In the training phase, Mask R-CNN requires a loss function to evaluate the training effectiveness and guide the training process in the right direction. The total loss function for Mask R-CNN is represented by the following equation:

$$L = L_{\text{box}} + L_{\text{cls}} + L_{\text{mask}},$$

(1)
where $L_{\text{box}}$ is the bounding box regression loss, $L_{\text{cls}}$ is the bounding box classification loss, and $L_{\text{mask}}$ is the mask loss [20].

### 2.2. Cascade Mask R-CNN

Despite the success of the two-stage architecture, the performance of water leakage segmentation remains suboptimal. In the Mask R-CNN framework, the region proposal network (RPN) generates regions of interest (RoIs) with an intersection-over-union (IoU) threshold of 0.5. However, this criterion may lead to prediction bias and ineffective segmentation of tunnel water leakage. To overcome this issue, Cascade Mask R-CNN adopts a cascade strategy. A more intuitive understanding can also be seen in Figure 3, where $H$ denotes the convolutional layer for extracting features, and $C$, $S$, and $B$ denote the classification, segmentation, and detection stages, respectively. The network framework utilizes RoI Align for feature extraction, which outperforms RoI pooling by eliminating the two quantization operations involved and employing bilinear interpolation to obtain image values at floating-point coordinates of pixel points [20].

![Cascade Mask R-CNN Network Framework](image)

Figure 3. Mask R-CNN and Cascade Mask R-CNN network framework.

Cascade Mask R-CNN introduces a split branch at each cascade stage to divide detection into three stages. The output of the previous stage serves as the input for the subsequent stage, and the three stages are supplied with RoIs having IoU thresholds of 0.5, 0.6, and 0.7, respectively, to enhance the quality of the output positive samples [25]. The accuracy of the target detector can be gradually improved by using a cascading detection strategy with multiple stages. By gradually increasing the IoU threshold, the model can screen more accurate positive samples at each stage, thus improving the final detection performance. Using different IoU thresholds can help the model adapt to targets of different sizes and difficulty. A lower IoU threshold (IoU = 0.5) can help the model to quickly localize the approximate location of the target, while a higher IoU threshold (IoU = 0.7) ensures that the model locates the target more accurately. Therefore, Cascade Mask R-CNN chooses to use a cascade detection strategy with IoU thresholds of 0.5, 0.6, and 0.7, aiming to improve
the accuracy of the target detector and achieve better detection performance in different target situations.

2.3. MDConv (Modulated Deformable Convolution)

The convolution method utilized in the baseline model maintains the same receptive field size across different locations, which results in an inability to automatically adjust the scale in regions of tunnel water leakage exhibiting various scales or deformations. This approach proves suboptimal for convolving at the edges of water leakage, given the irregular nature of such regions. To address this limitation, this paper employs deformable convolution instead of the original convolution method in the model, aiming to enhance the backbone network’s capability to extract water leakage features.

Deformable convolution builds upon the standard convolution operation by introducing an offset to the sampling position, thereby extending the convolution process to a broader range to accommodate irregular edge positions during training. By incorporating a learnable deformation parameter, deformable convolution enables adaptive adjustments to the sampling position and shape of the convolution kernel based on local features in the feature map. This adaptive mechanism allows for better adaptation to the irregular shape of tunnel water leakage, as well as potential occlusions such as wires or pipelines within the tunnel, and the complexity of the background. A comparison between standard convolution and deformable convolution is illustrated in Figure 4 [26].

![Figure 4. Illustration of the fixed receptive field in standard convolution (left) and the adaptive receptive field in deformable convolution (right).](image)

Figure 4a is the illustration of the feature extraction position under the $3 \times 3$ standard convolution, and Figure 4b is the illustration of the feature extraction position under the deformable convolution. From Figure 4b, we can visualize the change of deformable convolution in all directions, which can adaptively adjust the feature extraction position according to the measured image and is more suitable for the detection of water leakage in tunnels. Using the standard convolution and variable convolution of the feature extraction, results are calculated, and the output feature results are expressed as shown in Equations (1) and (2) [27]:

\[
y(p_0) = \sum_{k=1}^{K} w_k \cdot (p_0 + p_n + \Delta p_n) \cdot \Delta m_k ,
\]  

(2)
2.4. Backbone Network RegNetX

In the process of water leakage detection, there are many kinds of noise effects in the tunnel, and the ResNet-50 in the baseline model has a small receptive field, which leads to a small mapping area of the pixels of the feature map on the input image. This must be improved for the detection of such a complex environment as the tunnel [28]. In this paper, RegNetX is used to replace ResNet-50 in the baseline model as the feature extraction model, which can be adjusted to obtain a larger receptive field by adjusting different network structures through the number of layers and width of the network, and the network architecture is parameterized by investigating the network design space [29]. This model is faster in detection speed compared to the EfficientNets model. To analyze the quality of network design space, we mainly use the error empirical distribution function as shown in the following equation:

\[ F(e) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[e_i < e]}, \]  

where \( n \) represents the models and \( F \) is the proportion of models with errors less than \( e \).

2.5. Improved Backbone Network RegNetX-MDConv

The introduction of a mask cascading strategy in the Mask R-CNN baseline model is coupled with the integration of the RegNetX backbone network. Leveraging the larger receptive field area provided by RegNetX, MDConv is incorporated to adapt to water leakage areas at varying scales or deformations. This enhanced backbone network model is referred to as RegNetX-MDConv. With the improvement of the backbone network, RegNetX, the model can effectively capture a broader sensory field and extract features from the edge region of tunnel water leakage utilizing deformable convolution. The framework of this enhanced backbone network is depicted in Figure 5.

![RegNetX-MDConv network framework](image)

Figure 5. RegNetX-MDConv network framework.

The backbone network RegNetX is structured into three main components: stem, body, and head, as illustrated in Figure 5 [29]. The stem segment comprises 3 × 3 convolutional blocks, followed by batch normalization (BN) and rectified linear unit (ReLU) activation.

The body section is composed of four stages, each containing an S2 Block and multiple S1 blocks. In stages 2, 3, and 4, the group convolution of 3 × 3 is replaced with a 3 × 3
The backbone network RegNetX is structured into three main components: stem, body, and head, as illustrated in Figure 5 [29]. The stem segment comprises 3 × 3 convolutional blocks, followed by batch normalization (BN) and rectified linear unit (ReLU) activation.

The body section is composed of four stages, each containing an S2 Block and a fully connected (FC) layer [30]. The block structure within the body is depicted in Figure 6. The backbone network MRegNetX features a symmetrical block structure, which, when combined with deformable convolutions, greatly assists in extracting edge features from corresponding regions. The symmetrical block structure, when integrated with deformable convolutions, effectively extracts edge features from corresponding regions. Through the symmetrical structure, the network can exchange and propagate information in multiple directions, thus better capturing local features in the image. Meanwhile, deformable convolutions can dynamically adjust the shape and position of the convolutional kernel according to the deformation of specific regions, to adapt to the feature extraction requirements of different locations, thus more accurately capturing edge, texture, and other detailed information. This combination enables the network to achieve better results in tasks such as edge detection and feature extraction.

Figure 6. RegNetX-MDConv block structure.

3. Dataset Preprocessing

The dataset used in this paper comprises 383 tunnel water leakage pictures. There is considerable noise interference in the original tunnel water leakage pictures and the number of pictures is small, so the dataset is preprocessed, the dataset is expanded through preprocessing for noise reduction and dataset expansion, and the water leakage areas of tunnels are highlighted to improve the segmentation effect of the baseline model. The resolution of the tunnel leakage image used in this paper is 1115 × 1067.

3.1. Raw Data

The tunnel is poorly lit and there are other facilities (e.g., fasteners, hand holes, pipes, wires) inside the tunnel. The tunnel water leakage images taken in this environment will be affected by distortion, occlusion, and size change, which will influence the water leakage segmentation effect. Further processing of the raw data is required to reduce the dataset noise to highlight the tunnel water leakage characteristics. Moreover, the number of tunnel images used in this experiment is small, and the dataset needs to be expanded to obtain more positive samples for better baseline model detection.

3.2. Dataset Processing

For the water leakage in different areas of the tunnel, we use the subarea LabelMe annotation to label the water leakage areas more accurately. The original color water leakage pictures are grayed out, and the grayscale processing can then effectively reduce deformable convolution. The head segment includes a global average pooling (GAP) layer and a fully connected (FC) layer [30]. The block structure within the body is depicted in Figure 6. The backbone network MRegNetX features a symmetrical block structure, which, when combined with deformable convolutions, greatly assists in extracting edge features from corresponding regions. The symmetrical block structure, when integrated with deformable convolutions, effectively extracts edge features from corresponding regions. Through the symmetrical structure, the network can exchange and propagate information in multiple directions, thus better capturing local features in the image. Meanwhile, deformable convolutions can dynamically adjust the shape and position of the convolutional kernel according to the deformation of specific regions, to adapt to the feature extraction requirements of different locations, thus more accurately capturing edge, texture, and other detailed information. This combination enables the network to achieve better results in tasks such as edge detection and feature extraction.

Figure 6. RegNetX-MDConv block structure.
the parameters of the water leakage pictures without affecting the leakage segmentation and improve the robustness of the model. For the problem of fewer tunnel water leakage pictures, we rotate the dataset by 180 degrees, taking into account that the resolution of the pictures will affect the experiment. The number of tunnel water leakage pictures reaches 766 by expanding them. The specific operation is shown in Figure 7.

**Figure 7.** Tunnel leakage dataset preprocessing.

The detection performance (pixel segmentation) of water leakage images under different noise influences (such as fasteners, hand holes, pipes, and wires) varies greatly, making it difficult to meet the requirements of actual tunnel inspection projects. In this paper, 766 tunnel water leakage images are categorized into four main types of noise interference (as shown in Figure 8): Type I involves hand holes and pipe/wire interference, Type II involves leakage occurring at the edges of hand holes, Type III involves leakage around track wheel tracks, and Type IV involves leakage in simple backgrounds. By classifying the dataset, it is possible to analyze in more detail the factors affecting the detection performance of tunnel water leakage and summarize the overall effectiveness of tunnel water leakage detection [31].

Ablation experiments are performed based on the above dataset expansion and categorization to determine the enhancement effect on the baseline models Mask R-CNN and Cascade Mask R-CNN before and after preprocessing. As shown in Table 1, the AP, AP\(_{0.5}\), and AP\(_{0.75}\) of Mask R-CNN and Cascade Mask R-CNN before and after preprocessing are improved by 10.9% and 6.8%, 12.5% and 11.8%, and 15% and 12%, respectively.

**Table 1.** Average accuracy of Mask R-CNN and Cascade Mask R-CNN under DA (data augmentation).

<table>
<thead>
<tr>
<th>Model</th>
<th>AP</th>
<th>AP(_{0.5})</th>
<th>AP(_{0.75})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask R-CNN</td>
<td>0.300</td>
<td>0.564</td>
<td>0.287</td>
</tr>
<tr>
<td>Mask R-CNN+DA</td>
<td>0.409 (+10.9%)</td>
<td>0.689 (+12.5%)</td>
<td>0.437 (+15%)</td>
</tr>
<tr>
<td>Cascade Mask R-CNN</td>
<td>0.395</td>
<td>0.635</td>
<td>0.394</td>
</tr>
<tr>
<td>Cascade Mask R-CNN+DA</td>
<td>0.463 (+6.8%)</td>
<td>0.753 (+11.8%)</td>
<td>0.514 (+12%)</td>
</tr>
</tbody>
</table>

As shown in Figure 9, the pink curve and the blue curve represent the AP values before preprocessing for Mask R-CNN and Cascade Mask R-CNN, respectively. The green curve and the red curve represent the AP values under data augmentation (DA) for Mask R-CNN+DA and Cascade Mask R-CNN+DA, respectively. As can be seen from Table 1, after adding the cascade strategy, the AP, AP\(_{0.5}\), and AP\(_{0.75}\) are improved by 5.4%, 6.4%, and 7.7%, respectively, compared to the baseline model Mask R-CNN+DA. The specific AP data are shown in Figure 9.
Ablation experiments are performed based on the above dataset expansion and categorization to determine the enhancement effect on the baseline models Mask R-CNN and Cascade Mask R-CNN before and after preprocessing. As shown in Table 1, the AP, AP0.5, and AP0.75 of Mask R-CNN and Cascade Mask R-CNN before and after preprocessing are improved by 10.9% and 6.8%, 12.5% and 11.8%, and 15% and 12%, respectively.

As shown in Figure 9, the pink curve and the blue curve represent the AP values before preprocessing for Mask R-CNN and Cascade Mask R-CNN, respectively. The green curve and the red curve represent the AP values under data augmentation (DA) for Mask R-CNN+DA and Cascade Mask R-CNN+DA, respectively. As can be seen from Table 1, after adding the cascade strategy, the AP, AP0.5, and AP0.75 are improved by 5.4%, 6.4%, and 7.7%, respectively, compared to the baseline model Mask R-CNN+DA. The specific AP data are shown in Figure 9.

**Figure 8.** Tunnel water leakage dataset classification.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Train</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hand hole + pipeline/electric wire</td>
<td>134</td>
<td>30</td>
</tr>
<tr>
<td>Hand hole</td>
<td>172</td>
<td>24</td>
</tr>
<tr>
<td>Track wheel rail</td>
<td>108</td>
<td>18</td>
</tr>
<tr>
<td>Simple background</td>
<td>260</td>
<td>20</td>
</tr>
</tbody>
</table>

**Figure 9.** Average accuracy of Mask R-CNN and Cascade Mask R-CNN under DA (data augmentation).

4. Experiment Configuration

4.1. Configuration

All experiments in this study were conducted on a laboratory computer running the Ubuntu 20.04 operating system. The computer is equipped with an 11th Gen Intel(R)
Core(TM) i7-11700 processor running at 2.50 GHz (16 CPUs) and an NVIDIA GeForce RTX 3060 GPU. The experiments were implemented using MMDetection, a PyTorch-based target detection library developed by OpenMMLab, and the network framework was built upon this library. Throughout the experiments, a three-level cascade was employed. Each experiment involved training the detector for 12 epochs, starting with an initial learning rate of 0.005. The learning rate was reduced after the 8th and 11th epochs. The resolution of all water leakage images used in the experiments was set to 1115 $\times$ 1067 pixels.

4.2. Evaluation Indicators

In this paper, AP (average precision, which is the area enclosed by the P–R curve and the coordinate axis) is used as the evaluation index to evaluate the detection effect of different tunnel water leakage detection models, and the $AP_{0.5}$ and $AP_{0.75}$ involved in this paper are the AP values when IoU = 0.5 and IoU = 0.75, respectively. IoU is the intersection and concurrency ratio (as shown in Figure 10), and the calculation formula is as follows [32]:

$$\text{IoU} = \frac{S(B_p \cap B_{gt})}{S(B_p \cup B_{gt})} \quad (4)$$

![Figure 10. IoU under tunnel leakage detection.](image)

4.3. Model Training

We utilized the expanded dataset to train the enhanced model, Cascade-MRegNetX. The loss plot of the mask over 36 epochs is depicted in Figure 11. The loss value of the mask reflects the detection effect of tunnel water leakage. Based on the mask loss graph, it can be inferred that the improved model yields satisfactory experimental results. In this paper, we employed the Smooth L1 loss function in the model, which is a smooth loss function providing a robust measure of the disparity between the target region and the predicted region.

As shown in Figure 11, S0, S1, and S2 represent the mask loss curves corresponding to the first, second, and third stages, respectively. The Cascade-MRegNetX model was trained for 36 epochs. After the 12th epoch, the curves stabilized and the loss values were all below 0.2. Therefore, we selected 12 epochs for segmentation experiments in the paper. The learning rate was reduced at the 8th and 11th epochs to improve training effectiveness. The loss values corresponding to the three stages of the 13th epoch were 0.1832, 0.0992, and 0.0502, respectively.
4.3. Model Training
We utilized the expanded dataset to train the enhanced model, Cascade-MRegNetX. The loss plot of the mask over 36 epochs is depicted in Figure 11. The loss value of the mask reflects the detection effect of tunnel water leakage. Based on the mask loss graph, it can be inferred that the improved model yields satisfactory experimental results. In this paper, we employed the Smooth L1 loss function in the model, which is a smooth loss function providing a robust measure of the disparity between the target region and the predicted region.

3.6 9 1 2 1 5 1 8 2 1 2 4 2 7 3 0 3 3 6
0.0
0.1
0.2
0.3
0.4
0.5
0.1832
0.0992
0.0502
loss-mask
epoch
s0.loss-mask
s1.loss-mask
s2.loss-mask

Figure 11. The mask loss curve of Cascade-MRegNetX.

5. Segmentation Experiment
5.1. Replacing Backbone Networks
In this article, we use Mask R-CNN and Cascade Mask R-CNN as baseline models. Firstly, we use the expanded dataset to conduct segmentation experiments on five backbone networks: ResNet-50, GCNet, HRNet, RegNetX, and ResNeSt [33–35]. The comparison proves the effectiveness of the RegNetX backbone network in detecting tunnel water leakage. The experimental results are shown in Table 2 below.

<table>
<thead>
<tr>
<th>Model</th>
<th>Backbone</th>
<th>AP</th>
<th>AP_{0.5}</th>
<th>AP_{0.75}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask R-CNN+DA</td>
<td>ResNet</td>
<td>0.409</td>
<td>0.699</td>
<td>0.451</td>
</tr>
<tr>
<td></td>
<td>GCNet</td>
<td>0.422</td>
<td>0.698</td>
<td>0.445</td>
</tr>
<tr>
<td></td>
<td>HRNet</td>
<td>0.436</td>
<td>0.712</td>
<td>0.466</td>
</tr>
<tr>
<td></td>
<td>RegNetX</td>
<td>0.437</td>
<td>0.722</td>
<td>0.479</td>
</tr>
<tr>
<td>Cascade Mask</td>
<td>ResNet</td>
<td>0.463</td>
<td>0.753</td>
<td>0.514</td>
</tr>
<tr>
<td>R-CNN+DA</td>
<td>GCNet</td>
<td>0.466</td>
<td>0.732</td>
<td>0.510</td>
</tr>
<tr>
<td></td>
<td>HRNet</td>
<td>0.469</td>
<td>0.734</td>
<td>0.515</td>
</tr>
<tr>
<td></td>
<td>RegNetX</td>
<td>0.473</td>
<td>0.749</td>
<td>0.530</td>
</tr>
</tbody>
</table>

As can be seen from the comparison in Table 2, the AP of Cascade Mask R-CNN is better than Mask R-CNN under the same dataset and skeleton network, and the experiments show that better detection can be obtained by increasing the number of positive samples through the cascade strategy.

As shown in Figure 12, the AP curves for Cascade Mask R-CNN with five different backbone networks are displayed. The black curve represents the AP values of Mask R-CNN with data augmentation. As can be seen from the table, the replacement skeleton network has an improvement effect on AP, with RegNetX showing the greatest effect, its AP value improved by 2.8% and 1.0% compared with ResNet-50 of the two baseline models, respectively. The specific experimental results are shown in Figure 12.
5. Segmentation Experiment

5.1. Replacing Backbone Networks

In this article, we use Mask R-CNN and Cascade Mask R-CNN as baseline models. Firstly, we use the expanded dataset to conduct segmentation experiments on five backbone networks: ResNet-50, GCNet, HRNet, RegNetX, and ResNeSt [33–35]. The comparison proves the effectiveness of the RegNetX backbone network in detecting tunnel water leakage. The experimental results are shown in Table 2 below.

Table 2. Average accuracy of baseline model Mask R-CNN and Cascade Mask R-CNN with DA (data augmentation) under different backbone networks.

<table>
<thead>
<tr>
<th>Model</th>
<th>Backbone</th>
<th>AP</th>
<th>AP0.5</th>
<th>AP 0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask R-CNN+DA</td>
<td>ResNet</td>
<td>0.409</td>
<td>0.699</td>
<td>0.451</td>
</tr>
<tr>
<td></td>
<td>GCNet</td>
<td>0.422</td>
<td>0.698</td>
<td>0.445</td>
</tr>
<tr>
<td></td>
<td>HRNet</td>
<td>0.436</td>
<td>0.712</td>
<td>0.466</td>
</tr>
<tr>
<td></td>
<td>RegNetX</td>
<td>0.437</td>
<td>0.722</td>
<td>0.479</td>
</tr>
<tr>
<td>Cascade Mask R-CNN+DA</td>
<td>ResNet</td>
<td>0.463</td>
<td>0.753</td>
<td>0.514</td>
</tr>
<tr>
<td></td>
<td>GCNet</td>
<td>0.466</td>
<td>0.732</td>
<td>0.510</td>
</tr>
<tr>
<td></td>
<td>HRNet</td>
<td>0.469</td>
<td>0.734</td>
<td>0.515</td>
</tr>
<tr>
<td></td>
<td>RegNetX</td>
<td>0.473</td>
<td>0.749</td>
<td>0.530</td>
</tr>
</tbody>
</table>

As can be seen from the comparison in Table 2, the AP of Cascade Mask R-CNN is better than Mask R-CNN under the same dataset and skeleton network, and the experiments show that better detection can be obtained by increasing the number of positive samples through the cascade strategy.

As shown in Figure 12, the AP curves for Cascade Mask R-CNN with five different backbone networks are displayed. The black curve represents the AP values of Mask R-CNN with data augmentation. As can be seen from the table, the replacement skeleton network has an improvement effect on AP, with RegNetX showing the greatest effect, its AP value improved by 2.8% and 1.0% compared with ResNet-50 of the two baseline models, respectively. The specific experimental results are shown in Figure 12.

5.2. Experiments Related to the Model Cascade-MRegNetX

5.2.1. Transfer Learning

Migration learning refers to the use of a pre-trained model that has been trained on a large dataset to migrate to a target model using the corresponding structure and weights of that model. In this paper, the idea of migration learning is used to improve the segmentation performance of the model by using a pre-trained model provided by MMDetection that has been trained on big data and migrated to the baseline model in this paper. Considering the problem of small samples in the dataset used, the method of pre-trained model migration can also reduce the impact of small samples and obtain better training results.

Since replacing the baseline model backbone network destroys the original network weights, the pre-trained model needs to be fine-tuned. In this paper, we use the method of training specific layers and freezing other layers to fine-tune the model to update the network weights. When the pre-trained model for migration learning has been trained on large-scale data, the parameters already present strong generalization ability, so freezing these layers prevents overfitting on small-scale tunnel leakage datasets. The underlying network of the pre-trained model usually learns generalized features such as edges, textures, etc. Freezing these layers preserves their ability to extract features while adapting only the upper network to fit the model’s extraction of features in the tunnel leakage region. The 0, 1, 2, 3, and 4 stages of the model are frozen respectively and retrained to obtain new network weights. It is concluded from the freezing experiments that when the first stage is frozen, the model Cascade-MRegNetX can obtain a better migration effect, and the model AP obtained is 0.54. The specific experimental data are shown in Table 3.

As can be seen from Table 4, the AP is further improved after incorporating the transfer learning while other conditions remain unchanged. Among them, our model Cascade-MRegNetX obtains the highest AP value and achieves tunnel water leakage with high mask AP = 0.5400, AP0.5 = 0.7810, and AP0.75 = 0.6180. The specific experimental data are shown in Figure 13.
Table 3. Freezing the average accuracy of Cascade-MRegNetX with different number of layers after adding TL (transfer learning).

<table>
<thead>
<tr>
<th>Model</th>
<th>Frozen Stage</th>
<th>AP</th>
<th>AP&lt;sub&gt;0.5&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;0.75&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cascade-MRegNetX</td>
<td>0</td>
<td>0.5000</td>
<td>0.7440</td>
<td>0.552</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.5400</td>
<td>0.7810</td>
<td>0.6180</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.5030</td>
<td>0.7790</td>
<td>0.5820</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.4060</td>
<td>0.6780</td>
<td>0.4250</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.3730</td>
<td>0.6080</td>
<td>0.4190</td>
</tr>
</tbody>
</table>

Table 4. Average accuracy of Cascade Mask R-CNN under different backbone networks after adding TL (transfer learning).

<table>
<thead>
<tr>
<th>Model</th>
<th>Backbone</th>
<th>AP</th>
<th>AP&lt;sub&gt;0.5&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;0.75&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cascade Mask R-CNN+DA</td>
<td>ResNet</td>
<td>0.463</td>
<td>0.753</td>
<td>0.514</td>
</tr>
<tr>
<td></td>
<td>GCNet</td>
<td>0.521</td>
<td>0.762</td>
<td>0.510</td>
</tr>
<tr>
<td></td>
<td>HRNet</td>
<td>0.516</td>
<td>0.775</td>
<td>0.583</td>
</tr>
<tr>
<td></td>
<td>RegNetX</td>
<td>0.522</td>
<td>0.765</td>
<td>0.591</td>
</tr>
<tr>
<td></td>
<td>ResNetXt</td>
<td>0.502</td>
<td>0.728</td>
<td>0.560</td>
</tr>
<tr>
<td></td>
<td>MRegNetX</td>
<td>0.540</td>
<td>0.781</td>
<td>0.618</td>
</tr>
</tbody>
</table>

Figure 13. Average accuracy of different backbone networks after adding transfer learning.

As shown in Figure 13, the left graph depicts the comparison of AP between Cascade Mask R-CNN with different backbone networks after incorporating transfer learning and the baseline model. The right graph illustrates the comparison of AP between Cascade-MRegNetX after incorporating transfer learning and the baseline model. The red and green curves represent Mask R-CNN and Cascade Mask R-CNN, respectively. From Figure 13, it can be observed that after incorporating transfer learning, the AP values achieve good training results, showing a certain improvement from the beginning of training.

5.2.2. Ablation Experiments with the Model Cascade-MRegNetX

In this section, to further evaluate the robustness of the Cascade-MRegNetX network model for tunnel leakage detection, as well as to understand the key features and structure of the model in the decision-making process, ablation experiments are conducted to
isolate each component mask cascading strategy and the backbone network MRegNetX, and to clarify their respective contributions to the overall performance improvement.

From Table 5, it can be seen that the AP is further improved by adding the cascade strategy while other conditions remain unchanged. From the comparison between Mask-MRegNetX and Cascade-MRegNetX, it can be seen that the cascade strategy improves the AP, AP$_{0.5}$, and AP$_{0.75}$ of the model by 4.4%, 1.9%, and 6.9%, respectively. Other things being equal, AP is further improved by the addition of variable convolution in the backbone network RegNetX. From comparing Cascade-RegNetX with Cascade-MRegNetX, it is clear that the deformable convolution improves the AP, AP$_{0.5}$, and AP$_{0.75}$ of the model by 1.8%, 1.6%, and 2.7%, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>AP</th>
<th>AP$_{0.5}$</th>
<th>AP$_{0.75}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask-MRegNetX</td>
<td>0.496</td>
<td>0.762</td>
<td>0.549</td>
</tr>
<tr>
<td>Cascade-RegNetX</td>
<td>0.522</td>
<td>0.765</td>
<td>0.591</td>
</tr>
<tr>
<td>Cascade-MRegNetX</td>
<td>0.540</td>
<td>0.781</td>
<td>0.618</td>
</tr>
</tbody>
</table>

5.3. Analysis of Experimental Results
5.3.1. Segmentation Error Analysis

Detection results show four types of water leakage after improvement of the baseline model. As can be seen in Figure 14 below (the numbers on the left side represent the four types of tunnel water leakage mentioned above), the baseline model exhibits water leakage detection error as well as multiple repetitive detection. In this paper, the detection errors are categorized into three types, which are (1) E-A: erroneous detection, (2) E-B: missed detection, and (3) E-C: repeated detection. The specific error categories have been marked in the figure. The improved model will reduce the detection errors and increase the detection success rate.

Detection errors occur in the baseline model because the leakage area cannot be completely detected due to the occlusion of handholes, pipes, wires, and railroad wheel tracks. During the training process of the baseline model, the detection error occurs due to the limited feature extraction capability of the model. The improved model with deformable convolution enhances the feature extraction ability of the model, especially for the distinction of the edge part.

For class A and B errors, we can improve the quality of the samples through the method of dataset expansion to avoid the occurrence of false detection and improve the robustness of the model. By increasing the dataset expansion, the overfitting of the model to the training data can be mitigated and the performance of the model on unseen data improved. For class C errors, the main focus is to improve the detection accuracy through model improvement. In this paper, the feature extraction ability of the model is improved by adding cascading strategy and replacing the backbone network with MRegNetX.

The final experimental results are shown in the figure below. In Type 1 water leakage, the baseline model detects class A and C errors during the detection process, and the improved model corrects these errors and distinguishes between wires, pipes, and water leakage points. In Type 2 and Type 3 leaks, class A errors occurred during the baseline model detection process, and the improved model corrected these errors and achieved highly accurate mask prediction. In Type 4 leaks, the baseline model exhibited class B errors during detection, and the improved model identified the leakage areas. The improved model will thus reduce the detection errors and increase the detection success rate.
5.3.2. Segmentation Accuracy Analysis

By enhancing the baseline model Mask R-CNN through data augmentation, cascading strategy, transfer learning, and improving the backbone network, the Cascade-MRegNetX model has been proposed. This has led to improvements in the detection accuracy of four types of tunnel water leakages:

1. The detection accuracy of tunnel water leakage in simple backgrounds has increased from 64.0% to 98.4%;
2. The detection accuracy of tunnel water leakage around railway tracks has increased from 82.1% to 99.9%;
3. The detection accuracy of tunnel water leakage at the edge of handholes has increased from 44.3% to 96.8%, and optimization has been made to address the issue of one target having multiple detection boxes in the original baseline model;
4. The detection accuracy of tunnel water leakage around handholes and pipes/electrical wires has increased from 38.4% to 99.5%.

These are the improvements achieved by enhancing the Mask R-CNN model for the detection accuracy of four types of tunnel water leakages. Specific results are shown in Figure 15 (the numbers on the left side represent the four types of tunnel water leakages mentioned above).
6. Conclusions

This paper proposes a high-precision tunnel water leakage mask detection model based on Mask R-CNN, named the Cascade-MRegNetX model. In this model, RegNetX replaces the original backbone network ResNet-50 to achieve a larger receptive field and better network structure parameters. Additionally, MDConv is incorporated into the backbone network RegNetX to enhance the model’s feature extraction capability, particularly in irregular tunnel water leakage areas.

To achieve high-precision mask detection of tunnel water leakage, we adopt three optimization measures, including dataset classification and expansion, cascade strategy, and transfer learning. We use classification and expansion techniques to enhance the quantity and quality of the dataset, introducing a cascade strategy on the Mask R-CNN to provide higher-quality positive samples for detection, avoiding overfitting, and utilizing pre-trained models provided by MMdetection to improve the robustness of the model. Experimental results demonstrate that the proposed Cascade-MRegNetX detection model outperforms other mask detection models in detecting irregular regions such as tunnel water leakage. It achieves high-accuracy mask detection with AP = 0.54, AP$_{0.5}$ = 0.781, and AP$_{0.75}$ = 0.618. When compared with four different backbone networks (Cascade-ResNet-50, Cascade-GCNet, Cascade-HRNet, and Cascade-ResNeSt), our model exhibits superior accuracy.

Despite the satisfactory training results of this model, there are still some techniques that can enhance its performance. For instance, introducing a superior backbone network suitable for tunnel water leakage could improve segmentation accuracy. In the analysis of leakage results, distinguishing the types of leakage areas in the output results would facilitate the identification of leakage types in practical engineering scenarios.
Author Contributions: Conceptualization, X.X.; Methodology, W.W., X.X. and H.Y.; Formal analysis, W.W.; Investigation, W.W. and X.X.; Writing—original draft, W.W.; Writing—review & editing, X.X. and H.Y.; Supervision, X.X.; Funding acquisition, X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Jiangsu Province, China (BK20220502), Natural Science Foundation of China (No: 52308323, U1934209, 52090084), and Suzhou Innovation and Entrepreneurship Leading Talent Plan (No. ZXL2022488).

Data Availability Statement: The data provided in this work are available from the corresponding author.

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

References


Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.