DGN: A DWT-Guided Frequency-Spatial Dual-Domain Dehazing Network for Sewer Inspection Images

Gang Pan[1], Zhijie Sui[1], Zixia Xia[2], Chao Kang[3] and Di Sun[4]

[1] Tianjin University, Tianjin, China
[2] University of California, Irvine, California, USA
[3] University of Alberta, Alberta, Canada
[4] Tianjin University of Science and Technology, Tianjin, China

Abstract. Although spatial domain dehazing methods demonstrate effectiveness in removing dense haze, their limited capacity for holistic image restoration prompts exploration of frequency domain approaches. This paper presents a Discrete Wavelet Transform-guided Network (DGN) to address structure-aware dehazing in sewer images. The method enhances structure awareness through two key mechanisms: (1) a wavelet attention module that autonomously assigns weights to decomposed frequency components, with an emphasis on amplifying high-frequency features related to pipeline structures; (2) a contrastive regularization framework in the frequency domain designed to better preserve critical texture details during reconstruction. Experimental results demonstrate superior performance with 117 in mean square error (MSE), 28.27dB in peak signal-to-noise ratio (PSNR), and 0.9191 in structural similarity index measure (SSIM), achieved with 38.14MB parameters. Comparative studies highlight the complementary strengths of frequency- and spatial-domain approaches. the proposed DGN model demonstrates superior overall image restoration performance in full-image quality metrics, significantly improving hazy image scores in downstream tasks such as semantic segmentation, target localization, and classification. The proposed frequency- and spatial-domain fusion method provides an effective alternative solution for comprehensive sewer image restoration, particularly under complex haze distributions.

Keywords: Dehazing; DWT; domain fusion; sewer inspection

1 Introduction

Hazy images significantly impair the performance of visual tasks across various real-world applications, most notably in sewer pipeline inspection, where image clarity is critical for accurate defect detection and structural assessment. Consequently, developing effective haze removal algorithms tailored to pipeline inspection is of paramount importance. The Atmospheric Scattering Model (ASM)

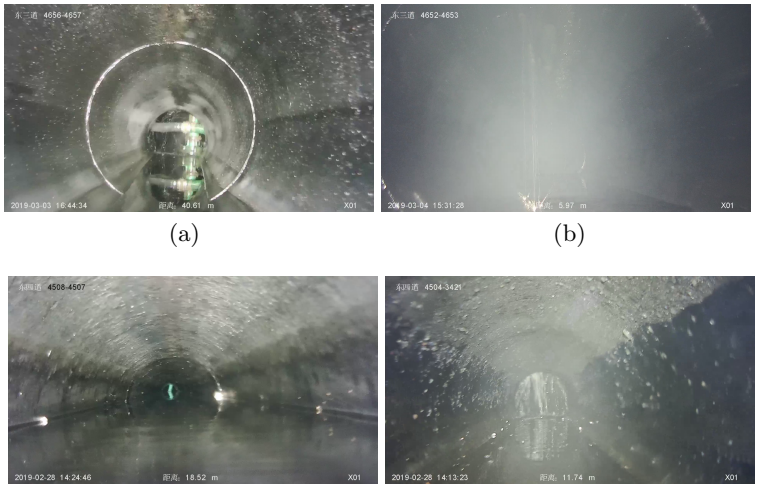


Fig. 1: Comparison between hazy area and haze-free area in sewer pipes of the same street. (a) Image of clear sewer in Dongsan St. (b) Image of hazy sewer in Dongsan St. (c) Image of clear sewer in Dongsi St. (d) Image of hazy sewer in Dongsi St.

offers a foundational framework for understanding haze-induced image degradation. It models hazy image formation as a combination of scene radiance and airlight, modulated by a transmission function that is strongly affected by particle scattering. Traditional image dehazing methods seek to recover the original scene radiance by estimating both the transmission map and atmospheric light.

Sewer pipelines are a vital component of urban infrastructure, responsible for the transport and management of wastewater. Regular inspections are necessary to maintain their structural integrity and operational efficiency. With the advancement of artificial intelligence (AI), automated pipeline inspection systems have been developed to enhance defect detection while reducing human labor reliance. However, the performance of AI models heavily rely on clean, high-quality images. Capturing such images in sewer environments presents several challenges: (1) high humidity and condensation within pipelines lead to substantial haze accumulation; (2) low-light conditions hinder effective image acquisition; and (3) reflections from water and airborne particulates further degrade image quality. As illustrated in Figure 1, sewer pipelines are typically humid environments where haze is prevalent. The presence of dense haze significantly impairs the ability to detect defects, such as leakage, as shown in Figure 1(b). In extreme cases, moisture may condense on the camera lens, further compromising image quality. Given these challenges, image dehazing has emerged as a critical preprocessing step to enhance image clarity and support the effectiveness of downstream AI-based inspection tasks.

Image dehazing has been extensively studied and can be broadly categorized into three approaches: (1) image enhancement-based methods, which improve contrast and visibility without explicitly modeling the haze formation process; (2) prior-based methods, which leverage physical priors such as the dark channel prior (DCP) or color attenuation prior (CAP) to estimate transmission maps and recover clear images; and (3) deep learning-based methods, which employ neural networks to learn complex haze removal mappings from large datasets.

To enhance dehazing performance, frequency-domain processing has been increasingly explored alongside traditional spatial-domain techniques. The Discrete Wavelet Transform (DWT) serves as a powerful tool in frequency-domain analysis, enabling the decomposition of images into multiple frequency components. Unlike conventional spatial-domain methods, DWT-based approaches facilitate the separation of low-frequency components—which capture the global structure of the image—from high-frequency components, which preserve edges and fine details such as sewer pipeline boundaries. This targeted decomposition makes DWT particularly well-suited for improving image dehazing in complex and visually challenging environments.

In our previous work [33], SANL-Net added attention to the two water boundary lines through multitask learning, achieving the goal of guiding the image dehazing process using structural information. As a spatial-domain dehazing algorithm, SANL-Net demonstrated the importance of structural cues; however, recent research has shifted toward leveraging multiscale frequency-domain information. In particular, high-frequency components are being used to enhance the network's focus on critical features such as water boundary lines. This motivated the development of a Frequency–Spatial Dual-Domain algorithm. In this study, a novel frequency-domain dehazing method based on the Discrete Wavelet Transform (DWT) is proposed for sewer pipeline inspection. The key contributions include: (1) the introduction of a wavelet attention mechanism to effectively decompose high-frequency signals and enhance edge and structural features, particularly water boundary lines; (2) the design of a multiscale dehazing network that balances spatial- and frequency-domain information to improve haze removal performance; and (3) the incorporation of frequency-domain contrastive regularization to align feature representations of hazy and dehazed images, thereby improving generalization. By leveraging frequency-domain characteristics to better preserve structural details, the proposed approach demonstrates superior performance compared to conventional methods, particularly in challenging sewer inspection scenarios.

2 Related Work

2.1 Image-dehazing networks

In the early stages, image-dehazing networks were developed based on ASM. Deep learning models were utilized to estimate parameters within ASM and then clean images could be obtained by computing ASM. DehazeNet [2], for example, estimated the transmission map t(x), and global atmospheric light A

was estimated empirically as a constant value. Consequently, the clean image was derived by solving the ASM equation. Li et al. [21] used a single parameter k(x) to represent t(x) and A both.

$$J(x) = K(x)I(x) - K(x) + b \tag{1}$$

$$K(x) = \frac{\frac{1}{t(x)}(I(x) - A) + (A - b)}{I(x) - 1}. \tag{2}$$

Thus, the proposed network was designed to predict k(x). Zhang and Patel [37] proposed to learn the transmission map t(x), the atmospheric light A, and the image-dehazing simultaneously. However, these approaches involve multiple steps, making them computationally tedious, and also may yield sub-optimal restoration performance due to inaccurate parameters.

Increasing attention is being directed toward end-to-end image dehazing networks, which employ deep learning models to directly estimate restored images without relying on explicit physical models. Researchers have explored various architectural innovations, including gated sub-network [3], dense feature fusion [7], Generative Adversarial Networks (GANs) [1, 28] and knowledge distillation techniques [16].Despite these advancements, the majority of existing methods primarily focus on enhancing network architectures, often overlooking the underlying image degradation process described by the Atmospheric Scattering Model (ASM).

Hu et al. [18] proposed a multitask learning model for joint image dehazing and depth estimation, based on the understanding that image degradation is directly influenced by scene depth. This approach requires both clean images and the corresponding depth maps as training labels. However, acquiring accurate depth information is often infeasible in sewer environments due to sensor limitations. To address this challenge, Xia et al. [33] leveraged the cylindrical geometry of sewer pipelines, suggesting that depth information can be inferred from the water borderlines formed between the inner wall of the pipeline and the residual water. Accordingly, a multitask framework was proposed to simultaneously learn image dehazing and borderline prediction. Experimental results demonstrated a strong correlation between dehazing quality and the accuracy of borderline prediction. Nonetheless, a key limitation remains—how to incorporate depth-related information for dehazing in a more generalizable manner without the need to introduce an additional prediction task.

2.2 Image restoration based on frequency domain

Image restoration networks can benefit significantly from frequency-domain information, which offers richer structural and textural cues compared to spatial-domain data alone. The Discrete Wavelet Transform (DWT), in particular, has been widely adopted to enhance reconstruction performance through various strategies. Guo et al. [9] applied DWT and its inverse (IDWT) before and after the network, respectively, allowing the entire forward propagation to operate within the frequency domain. Other studies have integrated multiple DWT

and IDWT operations throughout the network architecture to effectively fuse spatial- and frequency-domain features. For instance, Liu et al. [22] employed DWT in the encoder to extract structural information and used IDWT in the decoder to restore image details. Yang et al. [35] advanced this approach by decomposing frequency components and processing high- and low-frequency elements separately. Additionally, [30] developed a wavelet-based spatial attention module, which was embedded into the repeated blocks of the network to further exploit frequency-domain cues. To fully leverage frequency-domain information in image restoration, it is essential to strike a balance between spatial- and frequency-domain representations, as well as between high- and low-frequency components.

2.3 Dehazing objective function

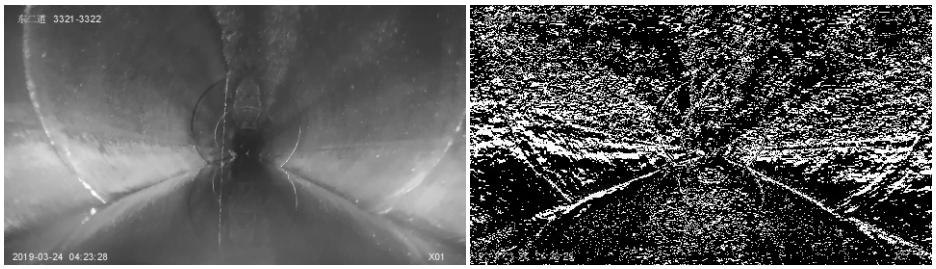
Most dehazing networks are trained using clean images as the sole supervisory signal without additional regularization [2, 21, 27, 8]. These methods rely on direct reconstruction loss between hazy and clean images, which often leads to suboptimal restoration, such as the loss of fine details or color distortion. To better exploit the information contained in clean images, several studies have introduced regularization strategies. Zhang and Patel [37], for instance, incorporated the transmission map t(x) of clean images as a regularized objective to provide depth-related guidance. Hong et al. [16] employed a teacher-student framework, where the teacher network extracted intermediate feature representations of clean images to guide the training of the student network.

While these approaches enhance the use of clean images as an upper bound for dehazing performance, limited attention has been given to incorporating hazy images as a lower bound. Wu et al. [32] addressed this gap by introducing contrastive learning to integrate both upper and lower bounds. However, their method relied on VGG-19 [29], to extract features solely from the spatial domain, neglecting frequency-domain characteristics that may also play a critical role in image restoration.

3 Methodology

3.1 Overview

As previously discussed, the Discrete Wavelet Transform (DWT) is employed to decompose images in the frequency domain, facilitating the extraction of fine-grained edge details that are essential for effective image dehazing. Unlike the traditional Fourier Transform, which captures only global frequency characteristics, DWT enables both frequency decomposition and spatial localization. This multi-resolution property makes DWT particularly suitable for tasks that require structural feature extraction, such as image dehazing. Among various wavelet families, the Haar wavelet is selected for its computational efficiency and effectiveness in capturing edge-like structures. Haar wavelets decompose an



(a) LL (b) LH

(c) HL (d) HH

Fig. 2: Four components in frequency domain.

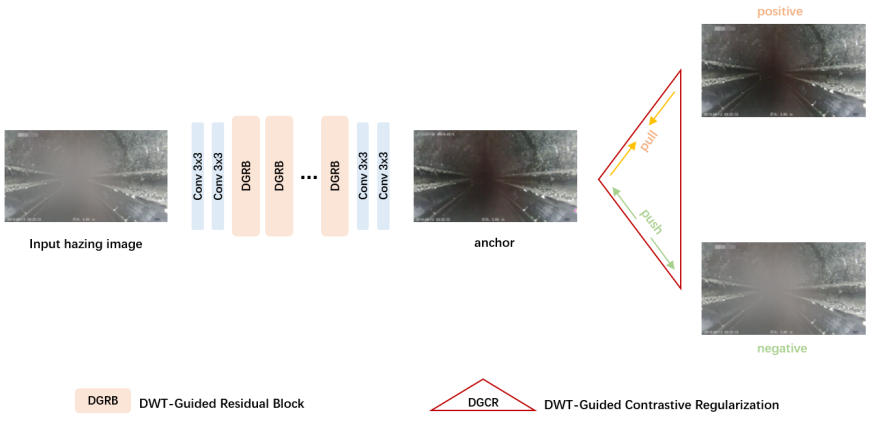


Fig. 3: The overall architecture of DGN.

input image X by computing the sum and difference of adjacent pixel values, resulting in four frequency components:

$$X_{LL}, X_{LH}, X_{HL}, X_{HH} = \mathrm{DWT}(X), \tag{3}$$

including LL, the low-frequency component, which represents the approximation coefficients of the original image; and LH, HL, and HH, the high-frequency components, which correspond to vertical, horizontal, and diagonal details respectively, as illustrated in Figure 2. The high-frequency components play a crucial role in capturing structural information, particularly highlighting the two water borderlines in sewer pipeline images. Since these water borderlines provide essential depth cues for image dehazing [33], this study introduces the DWT-Guided Network (DGN), designed to leverage high-frequency components to enhance structure-aware dehazing, as shown in Figure 3.

The proposed DGN architecture integrates frequency-domain information through two key components: DWT-Guided Residual Blocks (DGRBs) in the forward propagation and DWT-Guided Contrastive Regularization (DGCR) in the backpropagation. In the forward pass, the network architecture comprises two initial convolutional layers, followed by eleven DGRBs, and concludes with two additional convolutional layers. The initial and final convolutional layers serve as downsampling and upsampling modules, respectively, thereby reducing computational complexity and memory usage. The DGRBs are designed to enhance structural feature extraction by explicitly isolating high-frequency components from the input images, thereby improving the network's capacity to restore fine details that are often obscured or lost due to haze.

In the backpropagation process, DGCR introduces frequency-domain contrastive constraints to refine the dehazing performance. Given a restored image as the anchor, a clean image as the positive sample, and a hazy image as the negative sample, DGCR first transforms all three images into the frequency domain. It then optimizes the network by minimizing the feature distance between the anchor and the positive sample while maximizing the distance between the anchor and the negative sample. This ensures that the dehazed output is structurally and texturally aligned with a clean image while diverging from the hazy input.

Overall, DGRBs enhance attention to high-frequency components during forward propagation, thereby preserving fine structural details. Simultaneously, DGCR introduces contrastive regularization in the frequency domain, effectively guiding the network toward a more optimal and structurally consistent dehazing solution.

3.2 DWT-guided residual block

Figure 4 illustrates the technical details of DGRB. The primary objective of DGRB is to extract high-frequency components that convey structural information critical for image dehazing, while simultaneously preserving spatial-domain features. To achieve this, the DGRB adopts a dual-branch design: The upper part

is to extract high-frequency information; the lower part is to reserve spatial-domain information. At first, in order to extract high-frequency components from the frequency domain, DGRB uses DWT to decompose the input feature f into four components in the frequency domain, f_{LL}, f_{LH}, f_{HL} and f_{HH}, then abandons the low-frequency component and only processes the high-frequency components to enhance fine details:

$$f_H = Upsample(Concate(f_{LH}, f_{HL}, f_{HH})). \tag{4}$$

Haar wavelet is used in above DWT, and frequency components have the half size of the input image. Therefore, the upsample operation is carried out on the high frequency features obtained by concating. For the upsampled high-frequency feature f_H, For the upsampled high-frequency feature, we employ the Convolutional Block Attention Module[31], which sequentially applies the Channel Attention module followed by the Spatial Attention module. The Channel Attention module adaptively recalibrates feature responses by modeling inter-channel dependencies, allowing the network to focus on more informative channels while suppressing less relevant ones. The Spatial Attention module, on the other hand, enhances the feature representation by selectively emphasizing important spatial regions. By combining these two attention mechanisms, Convolutional Block Attention Module enables more effective feature refinement, improving the network's ability to capture fine-grained details in high-frequency components and enhancing the robustness of the dehazing process. The new output is multiplied with the original f_H. Finally, 3×3 convolutional layer reduces the number of channels, generating the refined frequency-domain feature f_f.

In the lower branch, to preserve spatial-domain information, the input feature f is processed through two 3×3 convolution layers, followed by a Efficient Multi-Scale Attention Module (EMA) [25] module. The Channel Attention (CA)[17] module enhances long-range spatial dependencies, ensuring that the restored image maintains structural consistency. EMA is similar to CA, but it divides the channels into groups for computation and utilizes both 1×1 and 3×3 convolutions to capture local and contextual information at different scales. By grouping channels, EMA reduces computational overhead while maintaining the effectiveness of attention mechanisms. The combination of 1×1 convolutions, which model cross-channel dependencies, and 3×3 convolutions, which capture spatial correlations, allows EMA to balance efficiency and expressiveness. This design enhances feature representation by selectively emphasizing important channels and spatial structures, making it particularly effective for tasks requiring fine-grained feature extraction. The output of this branch is denoted as f_s. The Gated Fusion Units are utilized to fuse the outputs from the frequency- and the spatial-domain. Two learnable matricesW_1, W_2 are obtained using the following formula, with α representing the weight of the learned frequency domain. α is generated using a sigmoid function to ensure that its value remains between 0 and 1. Meanwhile, 1-α represents the spatial domain weight. These two weights are used to adjust the feature outputs of their respective modules, where the weights indicate their contribution to the final output. Each residual

block has its own corresponding α, making the fusion process more flexible.

$$\alpha = sigmoid(W_1 f_f + W_2 f_s) \tag{5}$$

$$f'_f = \alpha \cdot f_f \tag{6}$$

$$f'_s = (1 - \alpha) \cdot f_s. \tag{7}$$

The final output of DGRB includes frequency domain information f'_f, spatial domain information f'_s and the feature map of the initial input f:

$$f_o = f + f'_f + f'_s. \tag{8}$$

This fusion strategy allows the network to recover image content while leveraging frequency-domain information to improve detail restoration. The inclusion of the initial input is inspired by the concept of residual learning [13], which mitigates the problem of vanishing gradients and facilitates the training of deep networks. By preserving the initial information, residual connections enable more effective feature propagation and improve the network's ability to capture fine details. Moreover, they allow the model to focus on learning the differences introduced by the transformation, leading to a more stable convergence and a better generalization.

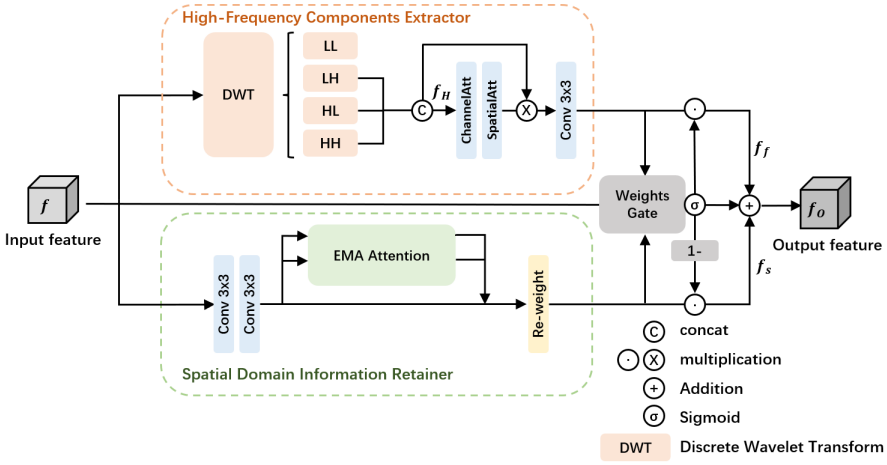


Fig. 4: The overall architecture of DGRB.

DGN consists of 11 DGRB blocks, each utilizing different dilation rates to extract multiscale features. The features captured at different layers are visualized in the figure 5, where brighter regions indicate areas of higher attention. It can be observed that the initial layers use low dilation rates to capture local details, while the middle layers adopt moderate dilation rates to extract large-scale structural information. In the final layers, low dilation rates are used again

to integrate all information. This multiscale strategy effectively preserves local details while capturing global features. The outputs of figure 5 are shown in 6.

3.3 DWT-guided contrasitve regularization

To enhance the network's convergence in the frequency domain, the DWT-Guided Contrastive Regularization (DGCR) module employs contrastive learning techniques [10, 24, 15, 14, 6]. The objective of DGCR is to minimize the feature distance between dehazed and haze-free images in the frequency domain while maximizing the distance between dehazed and hazy images, thereby improving dehazing effectiveness. DGCR involves two primary steps: constructing positive and negative feature pairs and identifying the latent feature space in which these pairs are compared. Three inputs are utilized: the output image of the network (referred to as the "anchor"), the corresponding clear image (the "positive"), and the hazy input image (the "negative"). Positive feature pairs are formed between the anchor and the positive sample, whereas negative pairs are constructed between the anchor and the negative sample. The Discrete Wavelet Transform (DWT) serves as the feature mapping function G, projecting all inputs into the frequency domain for subsequent contrastive comparisons. The objective function of the contrast regularization is formulated as follows:

$$\mathcal{L}_{DGCR} = \frac{\sum |G(\widetilde{I}) - G(I)|}{\sum |G(\widetilde{I}) - G(J)|}, \tag{9}$$

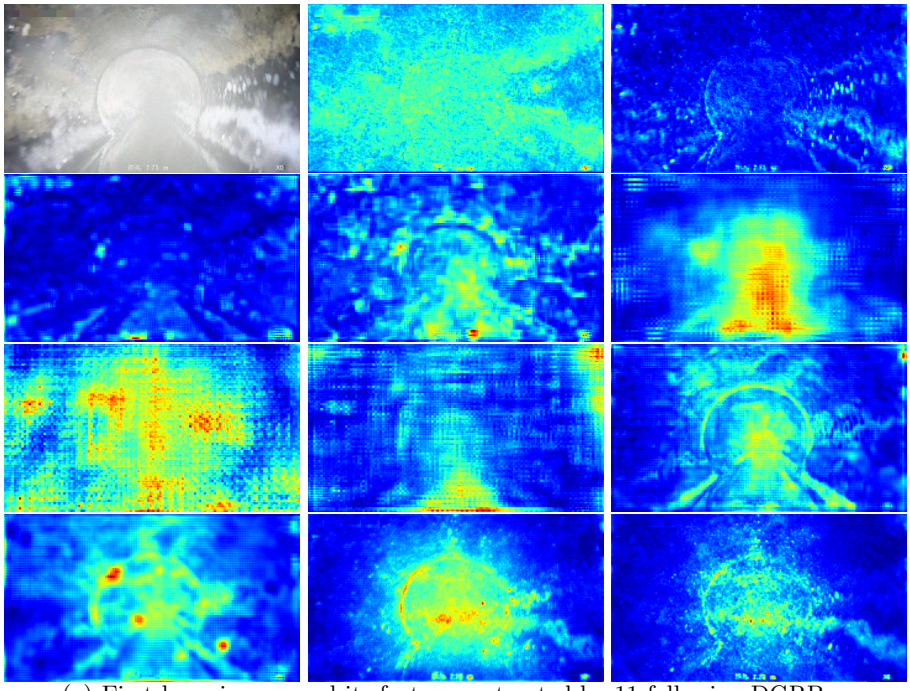
where G represents the function that converts the feature into the potential contrast space (i.e. DWT), $\widetilde{I}$ represents the image recovered by the network (i.e. anchor value), and I and J are the corresponding image without haze (i.e. positive value) and image with hazy (i.e. negative value). It can be seen that DGCR converts the feature space into the frequency domain, and not only reduces the difference between the restored image and the haze-free image in the frequency domain through forward regularization, but also uses the corresponding image with haze as a negative value to constrain the solution space in the frequency domain.

3.4 Loss functions

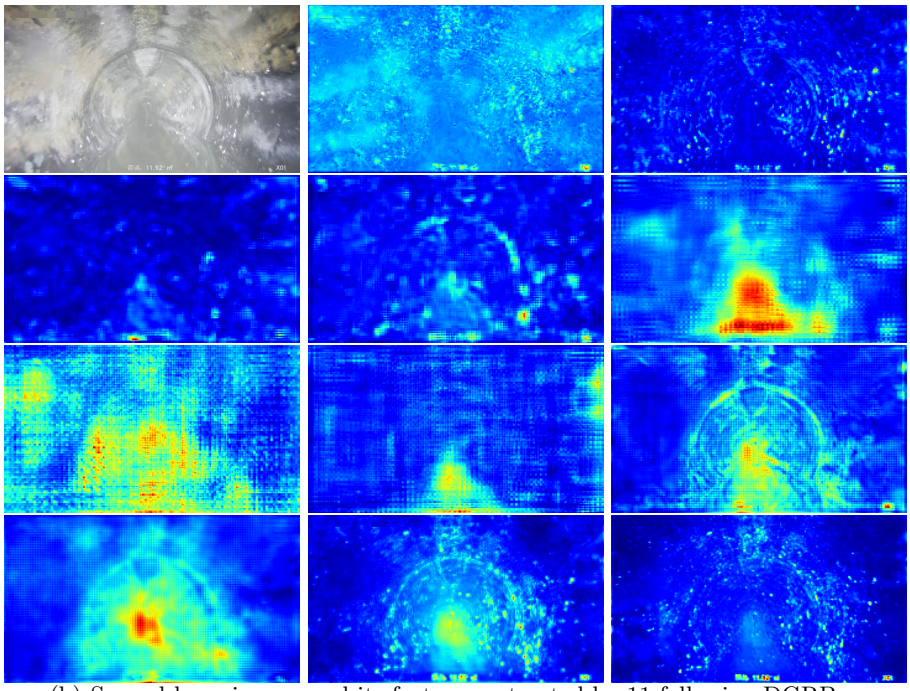
The loss function of DGN comprises two key components, namely DGCR and the reconstruction loss between the dehazed result and ground truth, therefore the complete formulation is expressed as:

$$\mathcal{L} = \sum |\widetilde{I} - I| + \mathcal{L}_{DGCR}, \tag{10}$$

where $\widetilde{I}$ and I are respectively the output and input of the network. Frequency domain algorithm not only uses DGRB to add common attention to spatial domain information and frequency domain information in the process of forward propagation, but also realizes the feature alignment of spatial domain and frequency domain in network training. Therefore, the algorithm greatly improves the convergence effect and enhances the quality of the haze removal image.

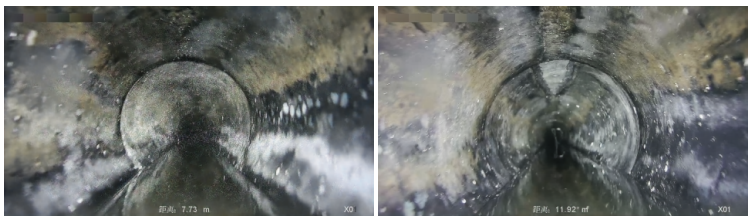


(a) First hazy image, and its features extracted by 11 following DGRBs.



(b) Second hazy image, and its features extracted by 11 following DGRBs.

Fig. 5: Features extracted by different DGRB modules in DGN.



(a) output of Fig.5(a) process (b) output of Fig.5(b) process

Fig. 6: outputs of figure 5(a) and (b)

4 Experiment

4.1 Training strategies

The DGN network was implemented using the PyTorch framework and trained on a GeForce RTX 4090D GPU. The Adam optimizer was employed for training, with a momentum parameter of 0.9 and an initial learning rate set to 5e-4. A polynomial (poly) learning rate decay strategy was adopted to adjust the learning rate throughout training. The network achieved optimal convergence at an iteration step size of 40,000.

4.2 Results

To evaluate image restoration quality, three widely adopted metrics are employed: Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM). These metrics provide complementary assessments of pixel-wise accuracy, perceptual fidelity, and structural preservation. The respective formulations of these metrics are as follows:

For image$I(x, y)$ and $\widetilde{I}(x, y)$:

$$\text{MSE} = \frac{1}{MN} \sum_{n=0}^{N} \sum_{m=0}^{M} |I(n, m) - \widetilde{I}(n, m)|^2 \tag{11}$$

$$\text{PSNR} = 10log_{10} \frac{peakval^2}{MSE}, \tag{12}$$

where peakval represents the maximum pixel intensity value in the image.

$$\text{SSIM(x, y)} = l(x, y)^{\alpha} c(x, y)^{\beta} s(x, y)^{\gamma}, \tag{13}$$

where l stands for luminance, c stands for contrast and s stands for structure, α, β, γ are constants.

A lower MSE value, and higher PSNR and SSIM scores, indicates smaller differences between the two images and better restoration quality. MSE and PSNR essentially measure pixel-wise differences across the images, whereas SSIM evaluates their similarity in terms of luminance, contrast, and structure, providing a more comprehensive reflection of image characteristics.

Table 1 presents a quantitative comparison of DGN on the test set. DCP [12], AOD-Net [21], GCANet [3], MSBDN-DFF [8], MPRNet [23], DGNL-Net [19], DCMP-Net [38] and SANL-Net [33] were used for comparison. As shown, DGN outperforms SANL-Net across all image quality metrics, including MSE, PSNR, and SSIM. In terms of parameter count, DGN contains more parameters than SANL-Net; however, it outperforms the parameter-intensive DCMP-Net in MSE and PSNR, while achieving comparable SSIM with significantly fewer parameters. This highlights the superior image restoration capabilities of DGN. Furthermore, during inference, DCMP-Net requires an average of 0.07290 seconds per image, whereas DGN only takes 0.02367 seconds—approximately 32.47% of the former—demonstrating its computational efficiency.

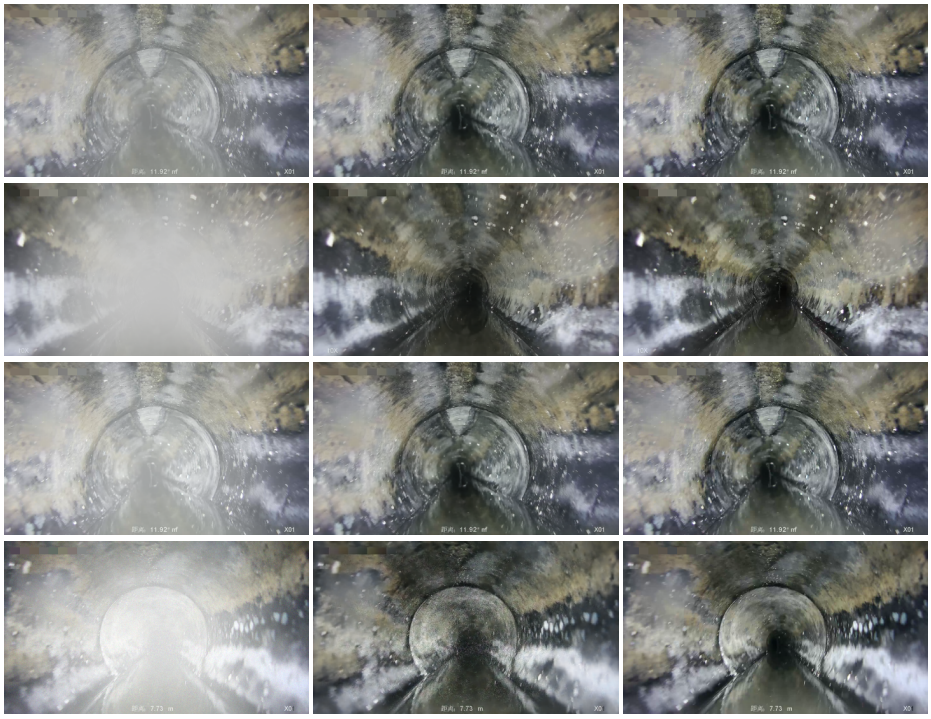
Table 1: Comparison of DGN and SOTA algorithms. Bold numbers represent first or second place.

Algorithms	**MSE**	**PSNR**	**SSIM**	**Parameters**
DCP(TPAMI'10)	3660	12.96	0.6843	
AOD-Net(ICCV'17)	2012	15.65	0.7458	0.01M
GCANet(WACV'19)	2696	14.53	0.7496	2.68M
MSBDN-DFF(CVPR'20)	172	27.10	0.9216	140.55M
MPRNet(CVPR'21)	655	20.72	0.8674	23.26M
DGNL-Net(TIP'21)	156	26.95	0.8921	15.40M
DCMP-Net(CVPR'24)	**143**	**28.13**	**0.9318**	199.65M
SANL-Net	147	27.28	0.8963	15.47M
DGN	**117**	**28.27**	**0.9191**	38.14M

Figure 7 shows the visualized results of DGN on the test set. It can be seen that the image recovered by DGN is more visually similar to the image without haze, both in structure and tone, such as the third and fourth lines. Figures 8 and 9 present the dehazing effect of DGN on the Pipe dataset and wild data, respectively, which were not used during training. It can be seen that the effect is still good and that the obstacles and details in the haze are clearer, demonstrating the good generalization performance of the DGN.

4.3 Ablation studies

This section presents ablation experiments conducted on the key components of DGN, specifically DGRBs and DGCR. In the DGRB ablation, the computation of frequency-domain features f_f was omitted, retaining only the spatial-domain



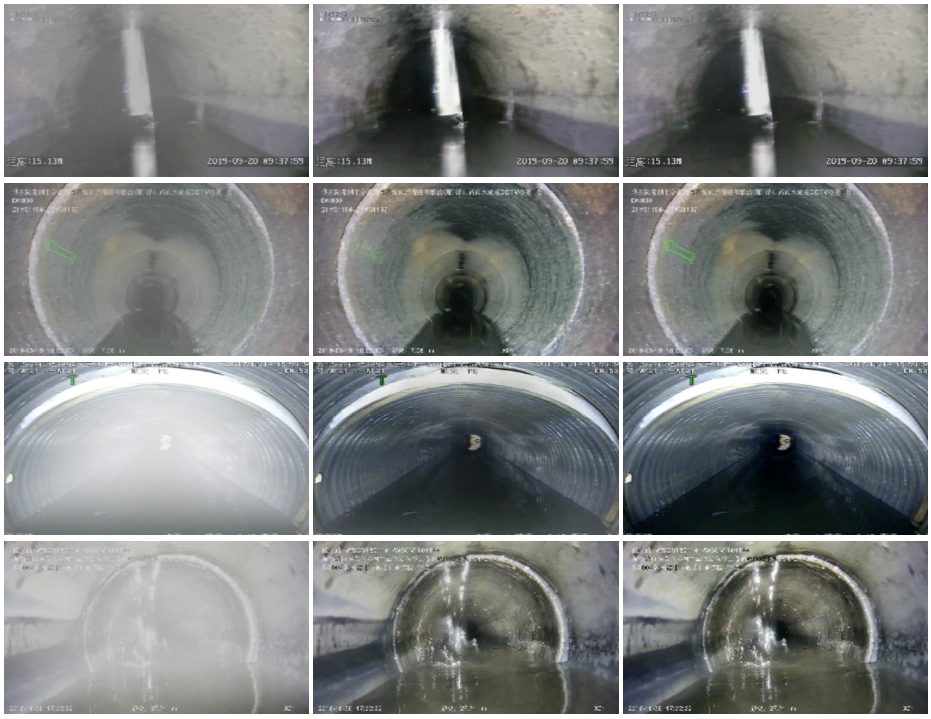
(a) input (b) output (c) groundtruth

Fig. 7: Results of DGN.

branch that calculates f_s. The results, summarized in Table 2 (where w/o denotes the removal of a specific component), indicate that both DGRB and DGCR contribute significantly to performance enhancement. Quantitative comparisons reveal that DGRB notably improves the SSIM value, while DGCR yields the most pronounced improvement in PSNR.

Table 2: Abalation studies

Algorithms	**MSE**	**PSNR**	**SSIM**
DGN w/o DGRB + DGCR	174	26.51	0.8934
DGN w/o DGCR	176	26.54	0.8960
DGN	146	27.43	0.9154



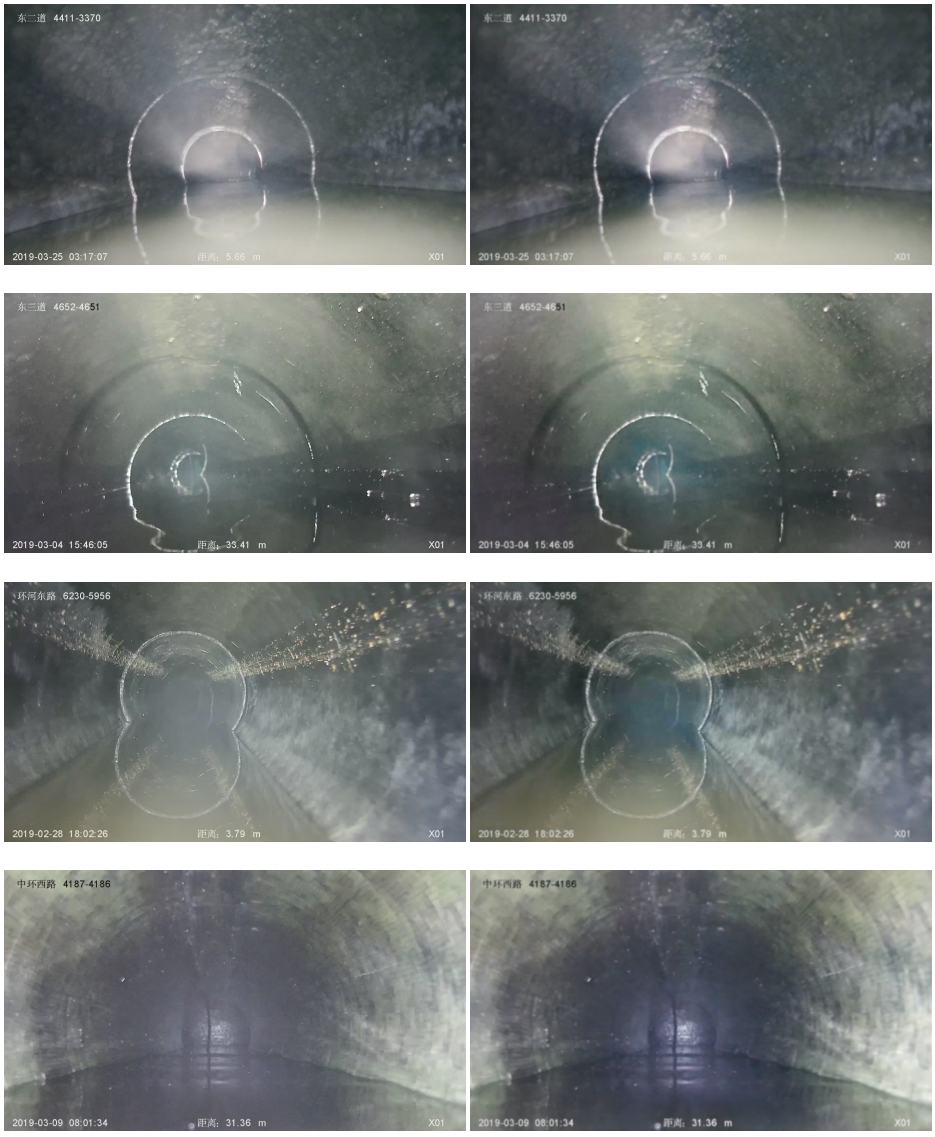
(a) input (b) output (c) groundtruth

Fig. 8: Results of DGN on Pipe dataset.

4.4 Frequency components in deep learning

In DGRB, the image is decomposed into four subcomponents LL, LH, HL and HH using DWT. The LL component functions as a low-resolution approximation of the image, capturing its low-frequency content, while the LH, HL, and HH components represent high-frequency information along the horizontal, vertical, and diagonal directions, respectively. These high-frequency subbands are essential for capturing edge structures, which are particularly critical in image dehazing tasks. DWT effectively addresses challenges commonly encountered in sewer pipeline imagery, such as large textureless regions and indistinct structural features. However, a key challenge lies in balancing spatial-domain and frequency-domain information and effectively integrating different frequency bands into the neural network architecture. To investigate this, three experimental configurations, A, B and C, were designed, as illustrated in Figure 10.

The input and output of these three experiments were paired images with and without haze. The overall architecture of the experiments follows a similar structure to that of the DGN, with modifications only made to the internal structure of the residual block. For all three experiments, only the first term of



(a) input (b) output

Fig. 9: Results of DGN on wild data.

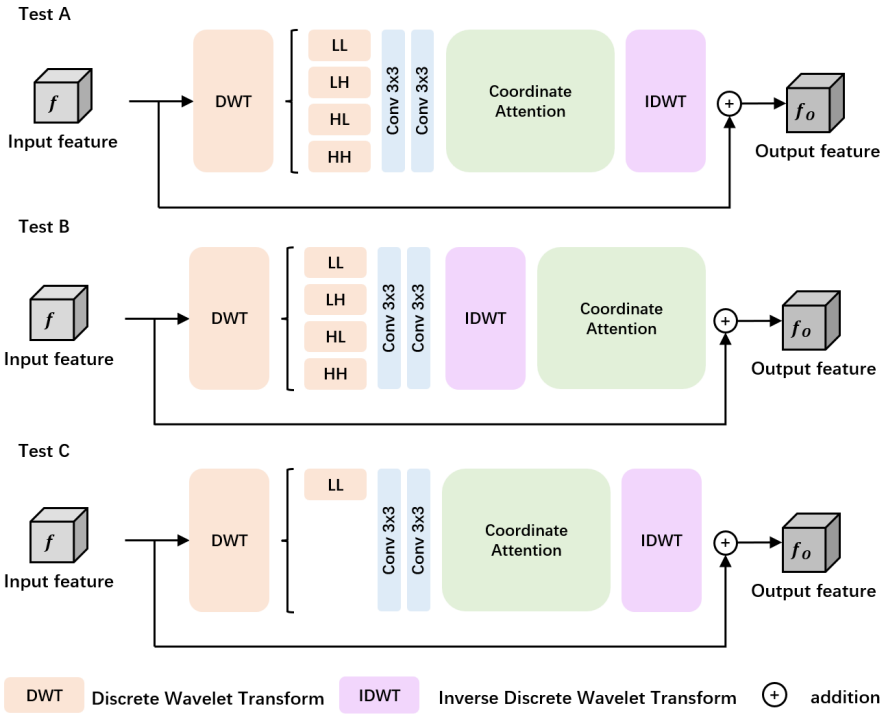


Fig. 10: Residual blocks in A, B, C.

formula (7) is used as the loss function to complete the optimization of network weight.

In experiment A, the image is converted to the frequency domain and then the convolution calculation is carried out, and then the coordinate attention module is used to strengthen the useful features. Finally, the feature layer obtained is restored to the spatial domain by IDWT. Compared with experiment A, experiment B applies the Coordinate Attention module in the spatial domain rather than in the frequency domain. Experiment C operates only on the low-frequency components in the frequency domain. These low-frequency components, along with the three untreated high-frequency components, are processed by IDWT to obtain the final output.

The quantitative results of these three experiments are presented in Table 3. Comparing Experiment A with Experiment B, it is evident that Experiment A only applies convolution operations to the features in the frequency domain, neglecting the valuable features in the spatial domain, which leads to poor performance. In contrast, Experiment B not only introduces attention mechanisms to the frequency-domain features but also retains the spatial-domain information, resulting in a significant improvement over Experiment A. When comparing Experiment A and Experiment C, it is observed that although the high-frequency

components contain essential edges and structural details, directly processing them alongside the low-frequency components in Experiment A yields worse results than processing only the low-frequency components, as in Experiment C. In summary, Experiment A shows the poorest dehazing performance, while Experiment B, which retains spatial-domain information and incorporates frequency-domain attention, demonstrates substantial improvements. Experiment C, which operates solely on the low-frequency components, also outperforms Experiment A. Thus, this experiment highlights the rationality and necessity of the DGRB module.

Table 3: The quantitative results of A, B and C experiments

Algorithm	**MSE**	**PSNR**	**SSIM**
A	260	24.59	0.8737
B	158	26.94	0.8942
C	146	26.10	0.8901

Based on the above findings, the final output of DGRB is composed of original feature f, frequency domain feature f_f and spatial domain feature f_s. This method has two advantages: (1) Convolution operation is carried out simultaneously in the frequency domain and the spatial domain, which preserves the structure and color characteristics of the two domains; (2) To ensure that different components, namely high frequency component and low frequency component are processed differently (low frequency component is the thumbnail most similar to the original image, and the original image has been processed in the calculation of f_s, so the calculation of f_f directly abandoned the low frequency component, so as to save computing space).

5 Defect detection improvements

Following the experimental settings in SANL-Net [33], this experiment evaluates the enhancement effects of DGN on three semantic tasks. The improvement brought by the proposed algorithm is assessed using three quantitative metrics: hazy value (HV), dehazed value (DV) and nondehazed value (UV). For the semantic segmentation and object localization tasks, the Pipe dataset [26] is used, while the Sewer-ML dataset [11] is employed for the image classification task. The performance of DGN is compared with that of SANL-Net to illustrate its effectiveness in enhancing semantic-level understanding.

5.1 Semantic segmentation

Table 4 shows that the results of the semantic segmentation experiments demonstrate a significant improvement in segmentation performance after applying the

DGN dehazing method. Due to the resolution differences between the defect dataset and the S2B dataset, two inference strategies are adopted to ensure the reliability of the results. The "w/o" strategy denotes direct inference on the original image resolution, while the "w/" strategy involves resizing the image to 480×272 to match the resolution of the S2B dataset prior to inference. These two strategies are also applied consistently in the subsequent object localization experiments.

The comparison between HV, DV, and UV reveals that the dehazed images consistently outperform the hazy images across all models (PipeUNet [26], FCN [36], and DeepLabv3+ [5]), achieving mIoU values closer to the clean images. For instance, the mIoU of the PipeUNet model for JO class improved from 47.83% in hazy images to 58.54% in dehazed images, approaching the clean image value of 62.25%. This trend is observed across other classes as well, suggesting that DGN effectively restores image details that are essential for precise segmentation.

Compared with SANL-Net (SANL-Net results can be referred to in our previous work due to space limitations), DGN achieves better performance in terms of MSE, PSNR, and SSIM values. As a result, it also provides greater improvements for semantic segmentation models, which can be observed from the semantic segmentation results. But in real pipeline image segmentation, SANL-Net has some advantages over DGN. Figure 11 illustrates the results of the defect JO detection in real images. After DGN restores the image, the semantic segmentation model can identify a larger defect area. However, compared to using SANL-Net, the detection accuracy decreases, and the probability of misclassification increases. For example, DGN causes FCN and DeepLabv3+ to incorrectly identify certain pixels. This is because DGN performs less effectively in processing the central region, whereas the defect JO in this image is precisely located in the center.

From the perspective of segmentation methods, DGN exhibits strong generalization and robustness, achieving notable improvements across different models. Regarding inference strategies, both the "w/o" (original resolution) and "w/" (resized to 480×272) strategies perform well on dehazed images. However, absolute scores for FCN and DeepLabv3+ under the "w/" strategy are slightly lower than those under "w/o", while hazy value (HV) scores are higher across all defect categories. This suggests that the "w/" strategy significantly affects the segmentation accuracy of hazy images but has minimal impact on dehazed or clean images. Notably, DGN effectively mitigates the accuracy drop caused by the "w/" strategy, narrowing the gap between it and the "w/o" approach. Overall, DGN significantly enhances semantic segmentation performance in various scenarios, making the segmentation of hazy images approach the level of ground-truth clean images.

5.2 Object localization

YOLOv5s, faster R-CNN and SSD used in the experiment are commonly used models for object localization [20]. In the experiments, as presented in Table 5, the mAP values also exhibit a clear improvement in localization accuracy after applying the DGN method. The dehazed images show a marked increase in mAP

Table 4: mIoU% of semantic segmentation

Model	Class	$\mathbf{hazy}^{HV}$	$\mathbf{dehazed}^{DV}$	$\mathbf{clean}^{UV}$
PipeUNet (w/o)	JO	$47.83^{14.42}$	$58.54^{10.71}$	$62.25^{3.71}$
	IL	$21.68^{26.45}$	$35.84^{14.16}$	$48.13^{12.29}$
	IN	$8.77^{20.02}$	$16.41^{7.64}$	$28.79^{12.38}$
	Mean	$26.09^{20.3}$	$36.93^{10.84}$	$46.39^{9.46}$
PipeUNet (w/)	JO	$29.08^{29.8}$	$54.55^{25.47}$	$58.88^{4.33}$
	IL	$10.67^{30.37}$	$38.54^{27.87}$	$41.04^{2.5}$
	IN	$5.53^{23.32}$	$21^{15.47}$	$28.85^{7.85}$
	Mean	$15.09^{27.83}$	$38.03^{22.94}$	$42.92^{4.89}$
FCN (w/o)	JO	$57.89^{5.47}$	$62.46^{4.57}$	$63.36^{0.9}$
	IL	$24.56^{21.45}$	$42.84^{18.28}$	$46.01^{3.17}$
	IN	$30.53^{20.41}$	$47.17^{16.64}$	$50.94^{3.77}$
	Mean	$37.66^{15.78}$	$50.82^{13.16}$	$53.44^{2.62}$
FCN (w/)	JO	$47.85^{12.15}$	$53.6^{5.75}$	$60^{6.4}$
	IL	$16.03^{21.11}$	$35.99^{19.96}$	$37.14^{1.15}$
	IN	$17.77^{34.02}$	$48.67^{30.9}$	$51.79^{3.12}$
	Mean	$27.22^{22.42}$	$46.09^{18.87}$	$49.64^{3.55}$
DeepLabv3+ (w/o)	JO	$60.58^{9.38}$	$66.82^{6.24}$	$69.96^{3.14}$
	IL	$16.54^{28.75}$	$39.79^{23.25}$	$45.29^{5.5}$
	IN	$42.12^{16.71}$	$55.47^{13.35}$	$58.83^{3.36}$
	Mean	$39.75^{18.28}$	$54.03^{14.28}$	58.03^{4}
DeepLabv3+ (w/)	JO	$40.24^{25.6}$	$56.61^{16.37}$	$65.84^{9.23}$
	IL	$9.49^{35.11}$	$44.56^{35.07}$	$44.6^{0.04}$
	IN	$34.58^{27.59}$	$56.43^{21.85}$	$62.17^{5.74}$
	Mean	$28.1^{29.43}$	$52.53^{24.43}$	57.53^{5}

compared to the hazy images. For example, YOLOv5s achieved a mAP of 74.16% for JO class in hazy images, which increased to 80.75% after dehazing, and was even higher to the performance of clean images at 80.04%. Similarly, Faster R-CNN and SSD models show improved localization accuracy in dehazed images across all classes. The substantial improvement in mAP for the IL and IN classes, particularly in YOLOv5s, suggests that DGN is highly effective in restoring spatial details that aid in accurate localization, thereby making it comparable to clean image performance. This reinforces the ability of DGN to reduce the adverse effects of haze on object detection tasks.

Compared with SANL-Net, DGN also achieves better performance. However, the improvement margin between these two networks is smaller in the object localization task compared to the semantic segmentation task.Figure 12 shows the effect of detecting defect IN on real images. When comparing different strategies, the w/o strategy provides a slight improvement in the interpretation score of hazy images. However, overall, both strategies have little impact on the absolute scores of the object localization task. Notably, DGN achieves a higher DV score under the w/ strategy, indicating that it performs well under both strategies, bringing hazy images with different scores to a similar level. When comparing different object localization models, under the w/o strategy, both networks achieve the best improvement on the SSD model. However, under the w/

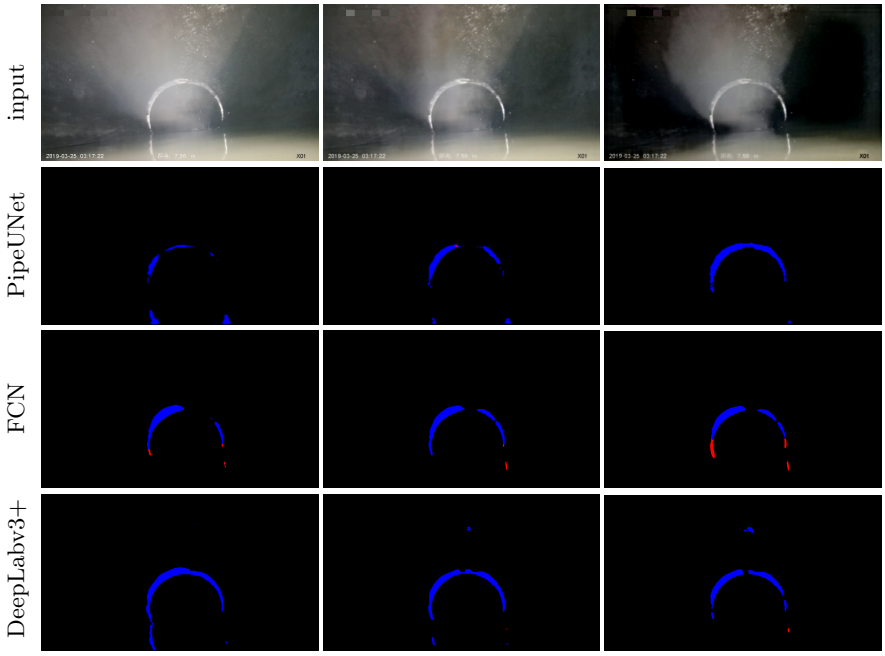


Fig. 11: Results of semantic segmentation experiments.

strategy, DGN provides the most significant improvement for YOLOv5s, while SANL-Net still achieves the best improvement for SSD.

5.3 Image classification

The image classification results, as shown in Table 6, further confirm the effectiveness of the DGN method in improving the performance of classification models. Due to the large number of categories in the experiment, the study is conducted using only the w/ strategy, and adopted GoogleNet InceptionV3 [4], ResNet-101 [34] and IDCNN&NDCNN [13] for classification. The F1 scores for the hazy images (HV) were generally lower than those for the dehazed images (DV), and the gap between the hazy and dehazed images was particularly notable in certain classes. For instance, GoogleNet InceptionV3 achieved an F1 score of 21.87% for RB class in hazy images, which improved to 29.37% after dehazing, approaching the clean image score of 31.11%. In the PF category, DGN achieves an HV improvement of 33.45, even surpassing the ground truth score, resulting in a negative UV value. Similarly, ResNet-101 and IDCNN and NDCNN models exhibited enhanced performance in the dehazed images, demonstrating that DGN plays a crucial role in restoring the features necessary for better classification. The consistent improvement across all three models further supports

Table 5: mAP% of object localization

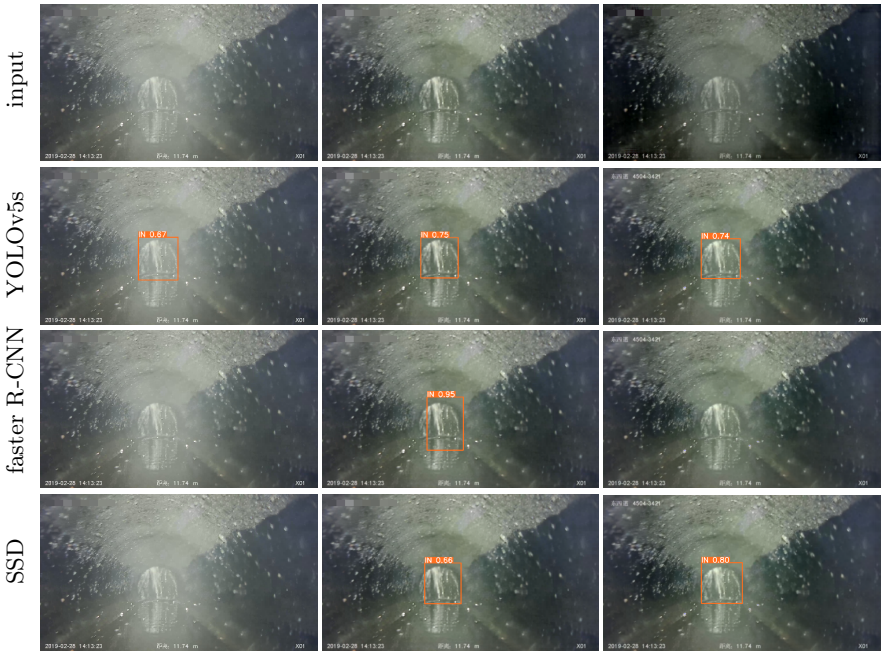
Model	**Class**	**hazy**^{HV}	**dehazed**^{DV}	**clean**^{UV}
YOLOv5s (w/o)	JO	$74.16^{5.88}$	$80.75^{6.59}$	$80.04^{-0.71}$
	IL	$42.74^{20.05}$	$60.15^{17.41}$	$62.79^{2.64}$
	IN	$43.88^{19.06}$	$61.47^{17.59}$	$62.94^{1.47}$
	Mean	53.59^{15}	$67.46^{13.87}$	$68.59^{1.13}$
YOLOv5s (w/)	JO	$61.49^{15.28}$	$75.24^{13.75}$	$76.77^{1.53}$
	IL	$44.48^{18.77}$	$63.19^{18.71}$	$63.25^{0.06}$
	IN	$38.24^{34.22}$	$66.52^{28.28}$	$72.46^{5.94}$
	Mean	$48.07^{22.76}$	$68.32^{20.25}$	$70.83^{2.51}$
Faster R-CNN (w/o)	JO	$62.65^{9.49}$	$68.83^{6.18}$	$72.14^{3.31}$
	IL	$28.8^{38.72}$	$54.37^{25.57}$	$67.52^{13.15}$
	IN	$33.16^{18.28}$	$37.84^{4.68}$	$51.44^{13.6}$
	Mean	$41.54^{22.16}$	$53.68^{12.14}$	$63.7^{10.02}$
Faster R-CNN (w/)	JO	$42.19^{4.16}$	$46.6^{4.41}$	$46.35^{-0.25}$
	IL	$29.52^{34.17}$	$60.4^{30.88}$	$63.69^{3.29}$
	IN	$21.44^{25.6}$	$41.87^{20.43}$	$47.04^{5.17}$
	Mean	$31.05^{21.31}$	$49.62^{18.57}$	$52.36^{2.74}$
SSD (w/o)	JO	$61.99^{8.11}$	$67.2^{5.21}$	$70.1^{2.9}$
	IL	$31.44^{22.79}$	$53.7^{22.26}$	$54.23^{0.53}$
	IN	$28.28^{23.23}$	$47.69^{19.41}$	$51.51^{3.82}$
	Mean	$40.57^{18.04}$	$56.2^{15.63}$	$58.61^{2.41}$
SSD (w/)	JO	$57.03^{13.32}$	$66.91^{9.88}$	$70.35^{3.44}$
	IL	$29.55^{25.03}$	$53.03^{23.48}$	$54.58^{1.55}$
	IN	$30.18^{21.01}$	$50.64^{20.46}$	$51.19^{0.55}$
	Mean	$38.92^{19.79}$	$56.86^{17.94}$	$58.71^{1.85}$

the conclusion that DGN enhances the recognition of key features by reducing haze-induced distortions in the images. However, it is worth noting that in certain categories, such as IN and OK, the accuracy gains from dehazing were minimal or even slightly negative. In these cases, the scores for hazy, dehazed, and clean images were quite similar, suggesting that haze has a limited impact on the classification of these particular defect types.

6 Conclusion

Haze is a common interference factor in CCTV-based sewer defect detection, substantially degrading image quality and reducing detection efficiency. This study investigates frequency-domain dehazing algorithms for sewer imagery and introduces DGN, a novel network that incorporates a wavelet attention mechanism and optimizes weights directly in the frequency domain. Although DGN has slightly more parameters than SANL-Net, our previous model, it achieves significantly improved performance in image restoration and defect detection.

This study also examines the role of the wavelet attention mechanism within DGN. The findings indicate that spatial-domain features remain essential for effective image restoration. The attention mechanism yields optimal performance



(a) hazy images (b) SANL-Net (c) DGN

Fig. 12: Results of object localization experiments.

Table 6: F1% of image classification

Model	GoogleNet InceptionV3			ResNet-101			IDCNN&NDCNN		
	Hazy^{HV}	**Dehazed**^{DV}	**Clean**^{UV}	**Hazy**^{HV}	**Dehazed**^{DV}	**Clean**^{UV}	**Hazy**^{HV}	**Dehazed**^{DV}	**Clean**^{UV}
RB	$21.87^{9.24}$	$29.37^{7.5}$	$31.11^{1.74}$	$23.1^{4.81}$	21.1^{-2}	$27.91^{6.81}$	$12.86^{9.57}$	$20.78^{7.92}$	$22.43^{1.65}$
OB	$60.43^{9.29}$	$70.06^{9.63}$	$69.72^{-0.34}$	$66.46^{8.77}$	$73.28^{6.82}$	$75.23^{1.95}$	$27.3^{34.82}$	$54.81^{27.51}$	$62.12^{7.31}$
PF	$30.47^{24.08}$	$63.92^{33.45}$	$54.55^{-9.37}$	$40.54^{18.92}$	$61.06^{20.52}$	$59.46^{-1.6}$	$10.34^{33.66}$	$26.79^{16.45}$	$44^{17.21}$
DE	$48.66^{18.01}$	$64.41^{15.75}$	$66.67^{2.26}$	$63.1^{17.71}$	$82.2^{19.1}$	$80.81^{-1.39}$	$5.29^{29.28}$	$40^{34.71}$	$34.57^{-5.43}$
FS	$63.94^{12.25}$	$73.93^{9.99}$	$76.19^{2.26}$	$64.41^{10.78}$	$75.61^{11.2}$	$75.19^{-0.42}$	$27.89^{39.64}$	$54.05^{26.16}$	$67.53^{13.48}$
IS	$7.78^{13.27}$	$14.36^{6.58}$	$21.05^{6.69}$	$15.85^{5.58}$	$19.94^{4.09}$	$21.43^{1.49}$	$1.18^{10.58}$	$5.97^{4.79}$	$11.76^{5.79}$
RO	$25.77^{11.27}$	$32.87^{7.1}$	$37.04^{4.17}$	$40.22^{-0.22}$	$37.77^{-2.45}$	$40^{2.23}$	$10.21^{6.81}$	$12.59^{2.38}$	$17.02^{4.43}$
IN	$39.01^{-3.45}$	$38.99^{-0.02}$	$35.56^{-3.43}$	$42.49^{-0.63}$	$35.82^{-6.67}$	$41.86^{6.04}$	$19.83^{-1.81}$	$19.53^{-0.3}$	$18.02^{-1.51}$
AF	$17.22^{16.11}$	$25.27^{8.05}$	$33.33^{8.06}$	$18.95^{13.48}$	$17.71^{-1.24}$	$32.43^{14.72}$	$6.15^{2.01}$	$11.93^{5.78}$	$8.16^{-3.77}$
BE	$43.48^{6.52}$	$50^{6.52}$	50^{0}	$55.99^{3.53}$	$52.25^{-3.74}$	$59.52^{7.27}$	$28.16^{7.74}$	$36.3^{8.14}$	$35.9^{-0.4}$
FO	$11.06^{6.33}$	$13.95^{2.89}$	$17.39^{3.44}$	$7.37^{13.68}$	$15.79^{8.42}$	$21.05^{5.26}$	$14.41^{-8.01}$	$6.36^{-8.05}$	$6.4^{0.04}$
GR	$28.9^{19.38}$	$40.79^{11.89}$	$48.28^{7.49}$	$42.58^{7.42}$	$45.71^{3.13}$	$50^{4.29}$	$11.16^{4.37}$	$14.09^{2.93}$	$15.53^{1.44}$
PH	$37.08^{9.07}$	$55.96^{18.88}$	$46.15^{-9.81}$	$42.65^{1.79}$	$47.16^{4.51}$	$44.44^{-2.72}$	$4.49^{9.8}$	$10.35^{5.86}$	$14.29^{3.94}$
OP	$19.57^{16.79}$	$26.45^{6.88}$	$36.36^{9.91}$	$44.83^{-8.47}$	$31.58^{-13.25}$	$36.36^{4.78}$	$5.26^{-0.13}$	$6.1^{0.84}$	$5.13^{-0.97}$
OK	$22.22^{-7.93}$	$20.31^{-1.91}$	$14.29^{-6.02}$	$65.93^{-5.93}$	$39.39^{-26.54}$	$60^{20.61}$	$2.19^{5.29}$	$5.86^{3.67}$	$7.48^{1.62}$
Mean	$31.83^{10.68}$	$41.38^{9.55}$	$42.51^{1.13}$	$42.3^{6.08}$	$43.76^{1.46}$	$48.38^{4.62}$	$12.45^{12.24}$	$21.7^{9.25}$	$24.69^{2.99}$

when low- and high-frequency components are processed separately. These observations are strongly supported by ablation experiments on DGRB. Furthermore, the paper introduces DGCR, a contrastive learning framework that enhances

network convergence by pulling positive feature pairs closer and pushing negative pairs apart in the frequency domain. Ablation studies on DGCR confirm its effectiveness. Improvements in defect detection are closely aligned with dehazing performance, with DGN outperforming SANL-Net overall. However, evaluations on real sewer images reveal nuanced differences: DGN achieves higher confidence in object localization tasks, whereas SANL-Net demonstrates greater accuracy in semantic segmentation.

References

1. Zahra Anvari and Vassilis Athitsos. Dehaze-glcgan: unpaired single image de-hazing via adversarial training. *arXiv preprint arXiv:2008.06632*, 2020.
2. Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016.
3. Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *IEEE winter conference on applications of computer vision*, pages 1375–1383. IEEE, 2019.
4. Kefan Chen, Hong Hu, Chaozhan Chen, Long Chen, and Caiying He. An intelligent sewer defect detection method based on convolutional neural network. In *2018 IEEE International conference on information and automation (ICIA)*, pages 1301–1306. IEEE, 2018.
5. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
6. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
7. Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.
8. Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2157–2167, 2020.
9. Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 104–113, 2017.
10. Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006.

11. Joakim Bruslund Haurum and Thomas B Moeslund. Sewer-ml: A multi-label sewer defect classification dataset and benchmark. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 13456–13467, 2021.
12. Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.
13. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
14. Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
15. Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020.
16. Ming Hong, Yuan Xie, Cuihua Li, and Yanyun Qu. Distilling image dehazing with heterogeneous task imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3462–3471, 2020.
17. Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13713–13722, June 2021.
18. Xiaowei Hu, Lei Zhu, Tianyu Wang, Chi-Wing Fu, and Pheng-Ann Heng. Single-image real-time rain removal based on depth-guided non-local features. *IEEE Transactions on Image Processing*, 30:1759–1770, 2021.
19. Xiaowei Hu, Lei Zhu, Tianyu Wang, Chi-Wing Fu, and Pheng-Ann Heng. Single-image real-time rain removal based on depth-guided non-local features. *IEEE Transactions on Image Processing*, 30:1759–1770, 2021.
20. Srinath Shiv Kumar, Mingzhu Wang, Dulcy M Abraham, Mohammad R Jahanshahi, Tom Iseley, and Jack CP Cheng. Deep learning–based automated detection of sewer defects in cctv videos. *Journal of Computing in Civil Engineering*, 34(1):04019047, 2020.
21. Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE international conference on computer vision*, pages 4770–4778, 2017.
22. Pengju Liu, Hongzhi Zhang, Wei Lian, and Wangmeng Zuo. Multi-level wavelet convolutional neural networks. *IEEE Access*, 7:74973–74985, 2019.
23. Armin Mehri, Parichehr B Ardakani, and Angel D Sappa. Mprnet: Multi-path residual network for lightweight image super resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2704–2713, 2021.
24. Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

25. Daliang Ouyang, Su He, Guozhong Zhang, Mingzhu Luo, Huaiyong Guo, Jian Zhan, and Zhijie Huang. Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
26. Gang Pan, Yaoxian Zheng, Shuai Guo, and Yaozhi Lv. Automatic sewer pipe defect semantic segmentation based on improved u-net. *Automation in Construction*, 119:103383, 2020.
27. Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11908–11915, 2020.
28. Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8160–8168, 2019.
29. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
30. Xibin Song, Dingfu Zhou, Wei Li, Haodong Ding, Yuchao Dai, and Liangjun Zhang. Wsamf-net: Wavelet spatial attention based multi-stream feedback network for single image dehazing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.
31. Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
32. Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10551–10560, 2021.
33. Zixia Xia, Shuai Guo, Di Sun, Yaozhi Lv, Honglie Li, and Gang Pan. Structure-aware dehazing of sewer inspection images based on monocular depth cues. *Computer-Aided Civil and Infrastructure Engineering*, 2022.
34. Qian Xie, Dawei Li, Jinxuan Xu, Zhenghao Yu, and Jun Wang. Automatic detection and classification of sewer defects via hierarchical deep learning. *IEEE Transactions on Automation Science and Engineering*, 16(4):1836–1847, 2019.
35. Mengping Yang, Zhe Wang, Ziqiu Chi, and Wenyi Feng. Wavegan: Frequency-aware gan for high-fidelity few-shot image generation. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022.
36. Xincong Yang, Heng Li, Yantao Yu, Xiaochun Luo, Ting Huang, and Xu Yang. Automatic pixel-level crack detection and measurement using fully convolutional network. *Computer-Aided Civil and Infrastructure Engineering*, 33(12):1090–1109, 2018.
37. He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2018.

38. Yafei Zhang, Shen Zhou, and Huafeng Li. Depth information assisted collaborative mutual promotion network for single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2024.