DOI: 10.1111/mice.12900

RESEARCH ARTICLE

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

Structure-aware dehazing of sewer inspection images based on monocular depth cues

Zixia Xia[1] | Shuai Guo[2] | Di Sun[3] | Yaozhi Lv[4] | Honglie Li[5] | Gang Pan[1]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China

[2]Department of Municipal Engineering, Hefei University of Technology, Anhui, China

[3]College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, China

[4]Key laboratory of Infrastructure Durability, Tianjin Municipal Engineering Design and Research Institute, Tianjin, China

[5]West-East Gas Pipeline Company, PipeChina, Shanghai, China

Correspondence
Gang Pan, College of Intelligence and Computing, Tianjin University, No. 135 Yaguan Road, Tianjin 300350, China.
Email: pangang@tju.edu.cn

Funding information
Natural Science Foundation of Tianjin, Grant/Award Number: 21JCYBJC00640; Key Research and Development Plan Support Program of Tianjin, Grant/Award Number: 20YFZCSN01080; Key Research and Development Program in Anhui Province, Grant/Award Number: 202104i07020012; National Cultural and Tourism Science and Technology Innovation Project of China, Grant/Award Number: 2021-97; University-Industry Collaborative Education Program, Grant/Award Number: 202102012011

Abstract

In sewer pipes, haze caused by the humid environment seriously impairs the quality of closed-circuit television (CCTV) images, which leads to poor performance of subsequent pipe defects detection. Meanwhile, the complexity of sewer images, such as steep depth change and extensive textureless regions, brings great challenges to the performance or application of general dehazing algorithms. Therefore, this study estimates sewer depth maps first with the help of the water–pipewall borderlines to produce the paired dehazing dataset. Then a structure-aware nonlocal network (SANL-Net) is proposed with the detected borderlines and the dehazing result as two supervisory signals. SANL-Net shows its superiority over other state-of-the-art approaches with 147 in mean square error (MSE), 27.28 in peak signal to noise ratio (PSNR), 0.8963 in structural similarity index measure (SSIM), and 15.47M in parameters. Also, the outstanding performance in real image dehazing implies the accuracy of depth estimation. Experimental results indicate that SANL-Net significantly improves the performance of defects detection tasks, such as an increase of 23.16% in mean intersection over union (mIoU) for semantic segmentation.

1 | INTRODUCTION

As essential components of civil infrastructure, underground sewer pipelines need regular inspections to ensure normal operation. Closed-circuit television (CCTV) carried by a robot vehicle is the most popular approach to capture pictures of the sewer internal condition. However, the manual inspection and condition assessment processes are low efficiency, error prone, and nonobjective due to the lack of sophisticated automatic defects detection

[Correction added on 22nd August 2022, after first online publication: The acknowledgement section and the multiplication signs were updated.]

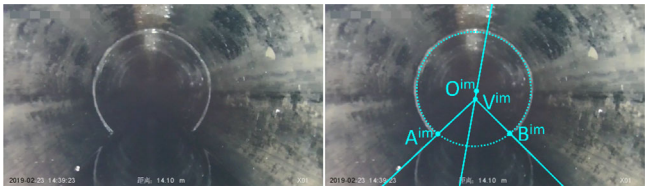


FIGURE 1 An image captured in a sewer environment with residual wastewater and based on the imaging model presented in Figure 2. Note that V_{im} is the vanishing point

tools under current practice. In many other industries, such as medical and construction fields, computer-aided systems have reached great successes (Hassanpour et al., 2019; Martins et al., 2020; Rafiei & Adeli, 2018). Therefore, in recent years, many studies including nonlearning methods (Iyer & Sinha, 2006) and machine learning methods (Cheng & Wang, 2018; Wang & Cheng, 2020) have appeared based on semantic segmentation, object localization, and image classification. All these methods need a large number of clear and clean sewer images for feature extraction (nonlearning-based methods) and for training (learning-based methods). However, in reality, CCTV typically produces hazy images due to the humid sewer environment. Although image dehazing in the computer vision field is a common task that has been studied for many years, there is a lack of published research on learning-based image dehazing in sewers probably due to its complexity and the lack of a public dataset.

The objective of image dehazing is to restore the clean counterpart of a single hazy image by removing the haze noise. The atmospheric scattering model (McCartney, 1976; Narasimhan & Nayar, 2001, 2002, 2003; Nayar & Narasimhan, 1999) is a classical description of haze degradation process

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{1}$$

$$t(x) = e^{-\beta d(x)} \tag{2}$$

where x indicates the pixel coordinate within the image, $I(x)$ indicates the hazy image pictured by the camera, $J(x)$ indicates the clean counterpart, $t(x)$ indicates the medium transmission map, A indicates the atmospheric light, β indicates the scattering coefficient of the atmosphere, and $d(x)$ indicates the image depth, which represents the distance from the object to the camera. Compared with other parameters, the image depth $d(x)$ is closely related to the hazy content. The far space in the image is full of thick haze while the near space suffers less from the haze. Figure 1 shows a clean image pictured in the sewer. It contains a wide range in depth, which changes substantially

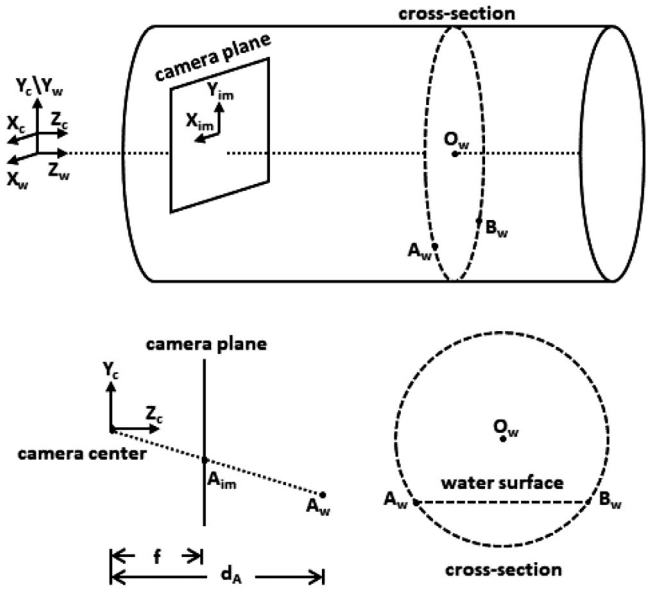


FIGURE 2 Scheme illustrating the back-projection of points. Figure 1 shows one image example

in one direction. What's more, similar colors and textures make it harder to obtain accurate depth information. Hence, the complexity in establishing the depth distribution makes the sewer image dehazing an engineering challenge.

Image dehazing is a common task that has been widely studied in the computer vision field and three types of methods have been proposed including image enhancement, prior-based algorithms, and learning-based algorithms. As for sewer image dehazing, image enhancement methods only improve the image clarity without considering the specific sewer environment and usually achieve unsatisfactory results (Guo et al., 2009, 2008; Kunzel et al., 2018). Prior-based algorithms have been widely applied but only one paper (Li et al., 2020) that used dark channel prior (DCP) to remove haze in sewer images has been found in literatures. However, prior-based algorithms are sensitive to the haze distribution, which results in their instability in real scenes. Learning-based algorithms based on convolution neural network (CNN) models show their superiority in many fields. In order to improve the performance of the defect detection task, Li et al. (2022) employed one previously proposed dehazing algorithm called gated context aggregation network (GCANet) (Chen et al., 2019) as a preprocessing step to sewer images. GCANet was proposed for general indoor and outdoor image dehazing and trained by public datasets. To train a network for sewer image dehazing, there is a need for a paired dataset including hazy images as inputs and clean counterparts as supervisory signals (Anvari & Athitsos, 2020). Although it is infeasible

for CCTV cameras to capture paired images, it is common to use the depth map to generate a hazy image via the atmospheric scattering model. However, the complexity of the sewer environment, such as darkness and extensive similar texture regions, limits the performance or application of most depth estimation algorithms. This paper first solves the depth estimation problem based on sewers' monocular cues.

This study introduces a learning-based method for sewer image dehazing based on an accurate estimation of the depth map of the sewer image. First, two borderlines serve as main cues to compute the geometrical relations of the sewer structure and then to construct the depth map. The depth map is then utilized to produce the paired dataset. Second, a learning-based method named structure-aware nonlocal network (SANL-Net) is proposed. SANL-Net not only takes advantage of CNN properties including robustness and accuracy but also takes the sewer structure into account by using two borderlines as an additional supervisory signal. Third, to assess the contribution of SANL-Net toward accuracy improvement of high-level vision tasks in sewer images, this study uses nine trained models to detect defects in hazy images, dehazed images produced by SANL-Net, and clean images for comparison.

2 | RELATED WORK

2.1 | Depth estimation

Depth estimation can be implemented via various visual cues including monocular, stereo, motion parallax, and focus cues. Unlike other approaches, methods based on monocular cues have the advantage of requiring no additional inputs. Those methods conduct depth estimation by utilizing texture cues, geometric cues, or the global structure. For example, Angot et al. (2010) proposed an initial depth map (IDM) bank for images with global structures (such as the image containing sky and ground or the image containing the street and buildings). Ge et al. (2017) used geometric cues including the vanishing line and vanishing point to indicate depth direction. Su et al. (2021) regarded the change from the vanishing point to other objects as two-dimensional Gaussian distribution.

In sewers, an early work (Cooper et al., 1998) located the vanishing point with the help of longitudinal mortar lines for brick sewers. In the case of modern sewers, which are mostly made of concrete, two water borderlines can play the same role as longitudinal mortar lines to find the vanishing point. Then, the depth map of a sewer image can change from the margin to vanishing point following a Gaussian distribution. To fine-tune the depth map, this work finds points on the same depth contour by using imaging characteristics that are motivated by an image stitching task proposed by Kunzel et al. (2018).

2.2 | Image dehazing

2.2.1 | Prior-based

There are plenty of conventional dehazing methods based on prior information, such as DCP (He et al., 2011), color attenuation prior (Zhu et al., 2015), and nonlocal color prior (Berman et al., 2016). Those methods concentrated on the estimation of the transmission map $t(x)$, which is used to recover the clean image via the atmospheric scattering model. However, prior-based approaches have rare applications in sewer images due to their instability and nonrobustness for practical applications (Li et al., 2020).

2.2.2 | Learning-based

Learning-based methods have two main types: intermediate manner and end-to-end manner. Intermediate manner represents approaches that combine conventional methods with CNN models, such as estimating transmission map $t(x)$ and/or atmospheric light A (Berman et al., 2016). However, those methods are not always accurate and cause suboptimal restoration performance. Instead of estimating intermediate parameters, AOD-Net (Li et al., 2017) first optimizes the end-to-end network to directly restore the clean image. Recent works, such as GCANet (Chen et al., 2019), MSBDN-DFF (Dong et al., 2020), and MPR-Net (Zamir et al., 2021), also conducted dehazing by the end-to-end way. Those methods avoid inaccuracy when recovering the final clean image but still remain unexpected noise-like artifacts. They concentrate on better network architectures and fail to consider the information of the image background. Therefore, another end-to-end work (Hu et al., 2021) proposed the depth-guided nonlocal network (DGNL-Net) to further improve the dehazing effects by considering scene depth information. It used the depth map to guide the whole network to remove haze. As for sewer images, borderlines can play the same role to guide CNN to realize a structure-aware dehazing.

2.3 | Multi-task learning (MTL)

MTL is a learning paradigm that learns multiple correlated tasks jointly. The knowledge contained in a task can be leveraged by other tasks. For example, joint training between semantic segmentation and object localization is

a typical multitask problem. Object localization provides the prior information for semantic segmentation (He et al., 2017; Pinheiro et al., 2016) while semantic segmentation offers global information and local details for object localization (Mao et al., 2017; Zhang, Qiao, et al., 2018). In road images, a joint lane segmentation and lane boundary detection (Zhang, Xu, et al., 2018) leverage a geometric relationship that the outline of the lane is exactly the lane boundary, while the area formed by the lane boundary is exactly the lane. In sewer images, wastewater borderlines are important cues for depth estimation to build the depth map, which plays a dominant role in image dehazing. Therefore, a joint network between water borderlines detection and image dehazing can leverage this geometric relationship.

3 | SEWER DEPTH ESTIMATION AND DATASET

3.1 | Sewer image depth estimation

As a cylindrical structure, the sewer has circular cross-sections, which are parallel with the camera plane. Image points on one cross-section have the same depth value and can be divided into two parts. The arc line part is comprised of sewer pipe points above the water surface and the straight line part represents points on the water surface. As shown in Figures 1 and 2, O_w, A_w, B_w are 3D points in the sewer, O_{im}, A_{im}, B_{im} are pixel points in the image.

$$O_wA_w = O_wB_w = r \tag{3}$$

$$V_wA_w = V_wB_w \tag{4}$$

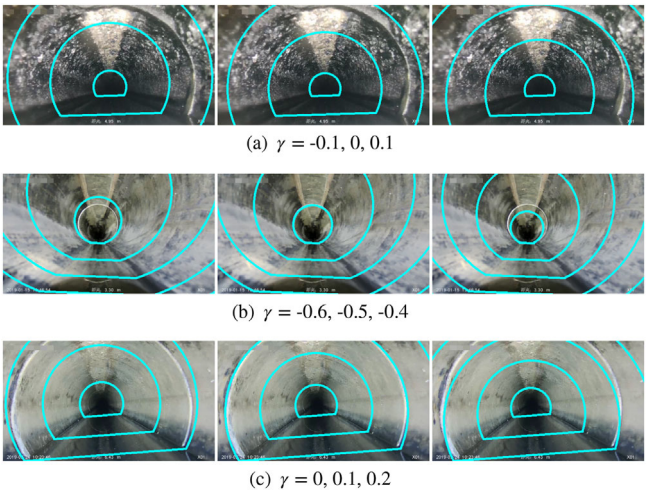
$$d_A = d_B \tag{5}$$

where r is the cross-section radius, d_A and d_B are depth values of the two points, V_w is the 3D intersection point between the optical axis and the cross-section. The intersection point between the optical axis and the camera plane is the vanishing point V_{im}. Under the pinhole camera model, the imaging characteristics can be given by:

$$\frac{O_wA_w}{O_{im}A_{im}} = \frac{V_wA_w}{V_{im}A_{im}} = \frac{O_wV_w}{O_{im}V_{im}} = \frac{O_wB_w}{O_{im}B_{im}} = \frac{V_wB_w}{V_{im}B_{im}} = \frac{d_A}{f} \tag{6}$$

Hence, it is easy to get:

$$O_{im}A_{im} = O_{im}B_{im} \tag{7}$$



(a) γ = -0.1, 0, 0.1

(b) γ = -0.6, -0.5, -0.4

(c) γ = 0, 0.1, 0.2

FIGURE 3 The fitting result for different γ. The middle image of each row indicates the best-fit

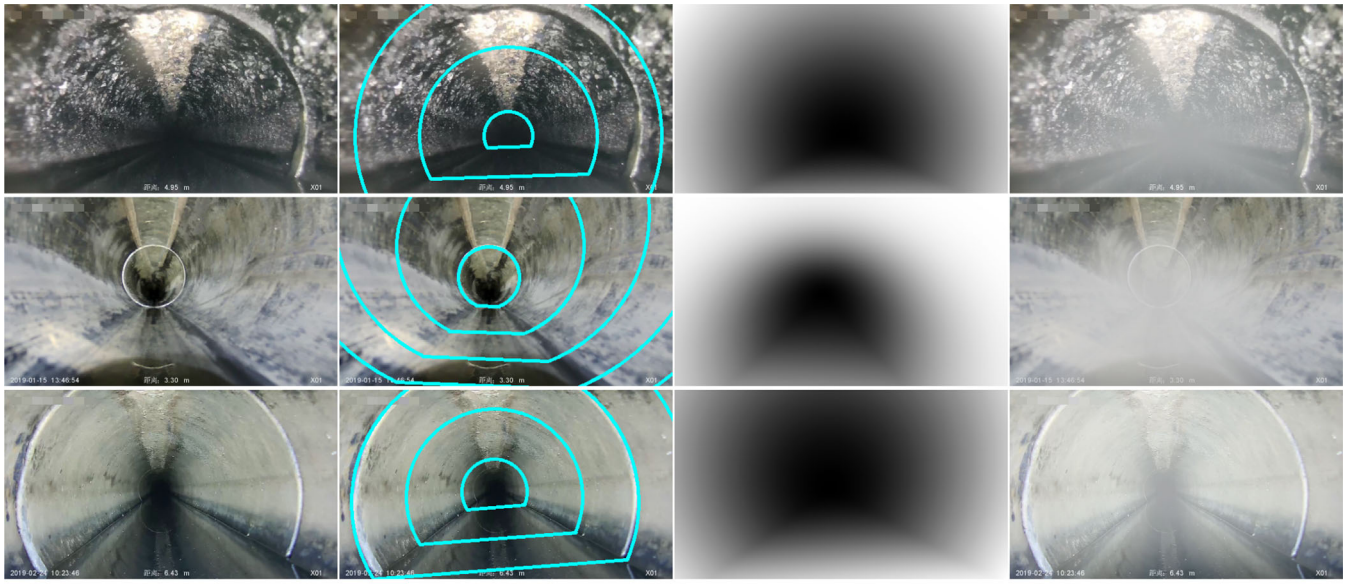
$$V_{im}A_{im} = V_{im}B_{im} \tag{8}$$

$$\frac{O_{im}V_{im}}{O_{im}A_{im}} = \frac{O_wV_w}{O_wA_w} \tag{9}$$

Equation (5) can be generalized to any pair of points on the same cross-section, so it can be deduced that the arc line is a circular arc. Equation (6) helps to locate one pair of points by finding two points that have the same distance to the vanishing point V_{im}. The circle center O_{im} is on the angular bisector of $\angle A_{im}V_{im}B_{im}$. Equation (7) helps to locate the exact position of the center. Specifying $\gamma = \frac{O_wV_w}{O_wA_w}$, the center can be identified by a known γ. Positive γ means that the center is above the vanishing point. Zero γ indicates the center and the vanishing point are the same. So depth contour for this pair of points is identified, which can be generalized to any pair of points on the two wastewater borderlines.

3.1.1 | Construct the depth map

This work chooses 19 alternative values in [−0.9, 0.9] with the interval of 0.1 to find the best-fitting γ for every image. Figure 3 shows the fitting effects of some values. Two factors determine the best-fitting γ: overlap degree between circle joint and depth contour and the identified γ from adjacent frames. Then the depth map is constructed with depth contours. The depth contour near the vanishing point has the maximum depth value. And depth values for depth contours take on a Gaussian distribution (Su et al., 2021). Figure 4 shows some examples.



(a) clean image (b) depth contours (c) depth map (d) hazy image

FIGURE 4 Some examples of the generation process of the hazy image

3.2 | The sewer-with-2-borderlines (S2B) dataset

Tianjin Municipal Engineering Design and Research Institute provides the CCTV videos, which were used for the pipe defect detection. Experimental images are acquired by capturing clean video frames, which come from different CCTV devices, regions, and cities. Then, wastewater borderlines are labeled manually to construct depth maps. As a real hazy photo is always a distant view, it means the vanishing point should be in the image center. Only clean frames whose intersection point between two borderlines is in the image center are remained. These remaining images reach 1090 and meet the condition that they cover various pipe materials and pipe diameters. Figure 6 shows their distribution in γ. It can be seen that the CCTV camera mostly locates in the cross-section center.

These clean images and their depth maps generate synthesized hazy images according to the atmospheric scattering model. Two sets of parameters (0.6, 0.8, 1), (1, 2, 3) are used separately for the atmosphere light and the scattering coefficient to simulate different degrees of haze. The comparison between simulation results and actual hazy images is shown in Figure 5. Synthesized images generated by these two sets of parameters can efficiently cover various actual hazy images and also generate some extreme conditions, which can help to enrich the dataset. Data augmentation (i.e., flip, noise) is also randomly involved to improve networks performance. Eventually, the S2B dataset contains 11,426 training images (captured at 22 streets) and 1314 validation images (captured at four other different streets), which are all resized to 480 × 272. The resolution is sufficient for depth estimation, without requiring a large amount of computation. Figure 4 shows some examples.

4 | DEHAZING METHODOLOGY

For image dehazing, high-level semantic information and low-level spatial information are all demanded in a balanced way (Zamir et al., 2021). As shown in Figure 7, SANL-Net contains three parts. The Semantic Net acts as an encoder–decoder architecture to learn the broad contextual information with large receptive fields. In addition to the high-level semantic information, this subnetwork also fully considers the atmospheric scattering model as the borderlines are important cues to estimate the depth map. The Spatial Net implements pixel-to-pixel correspondence from the input image to output image on the same resolution. The aim of the subnetwork is to preserve the desired fine texture and low-level spatial details in the final output image. Ultimately, the structure-aware nonlocal (SANL) module is applied to combine the local position information and the high-level semantic information by nonlocal operations to catch long-range dependencies.

4.1 | Semantic Net

The Semantic Net is composed of two main parts: an encoder–decoder network and deep Hough transform (DHT) layers. The encoder–decoder network usees 10

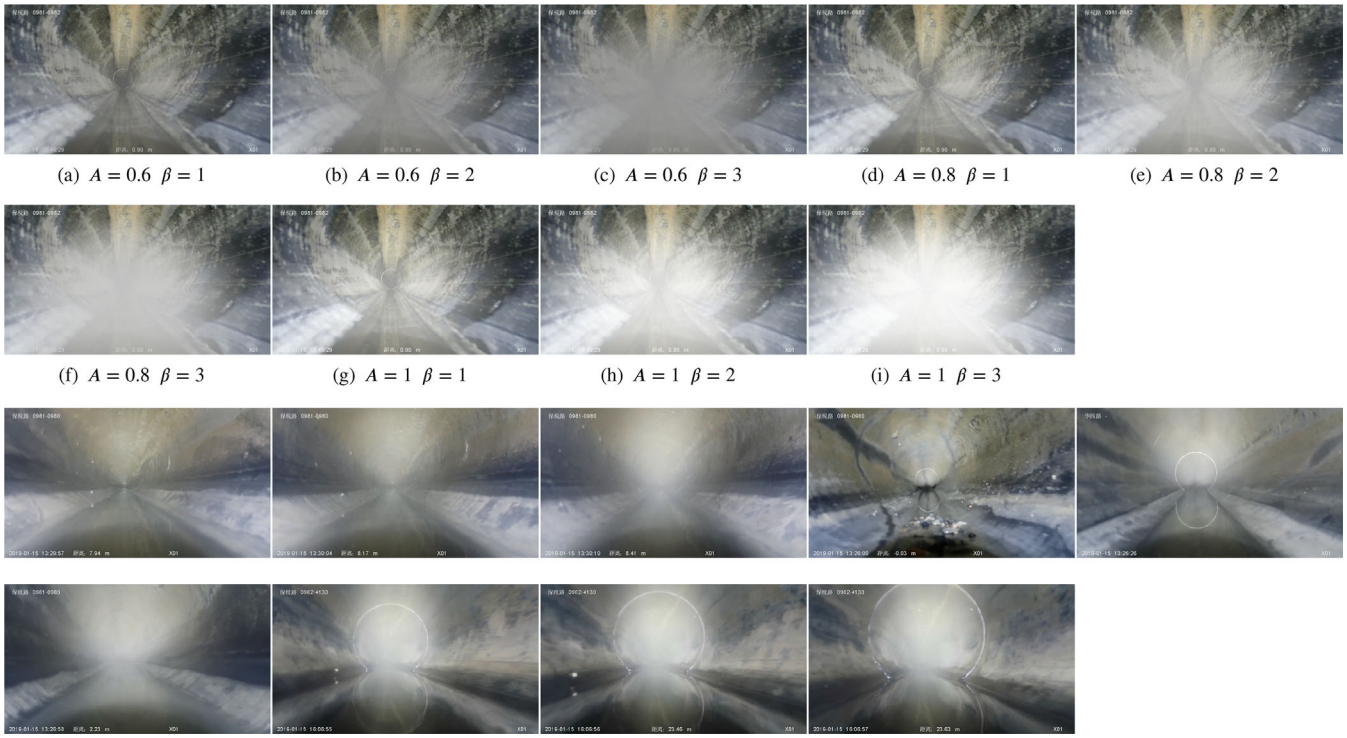


FIGURE 5 Comparison between synthesized hazy images and actual hazy images. The first two rows are synthesized hazy images and the last two rows are actual hazy images

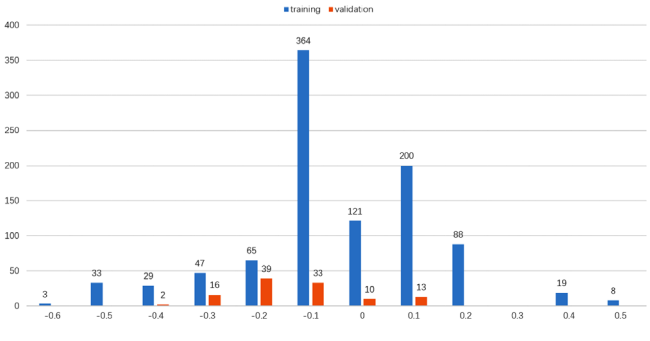


FIGURE 6 Distribution in γ

blocks to extract features. Every block contains a convolution layer, a group normalization (Wu & He, 2018), and a scaled exponential linear unit (SELU) (Klambauer et al., 2017). Ending blocks are connected by skip connections to retain information. The group numbers of group normalization of these blocks are all set to 32 and Table 1 shows convolution details. Finally, the output of the last block contacts with the output of Spatial Net in the SANL module.

After the encoder–decoder network, the four layers before the skip connections in the decoder are transformed to four feature maps at the size of $C \times \Theta \times R$ in parametric space (Hough space) by DHT layers (Zhao et al., 2021).

$$\Theta = \frac{\pi}{\Delta\theta} \tag{10}$$

$$R = \frac{\sqrt{W^2 + H^2}}{\Delta r} \tag{11}$$

where $\Delta\theta$ and Δr are the quantization interval for angle θ and radius r, W and H are, respectively, the image width and height, and C indicates the channel size. In this work, $C = 32$, $\Delta\theta = \frac{\pi}{100}$, $\Delta r = \sqrt{2}$. Then these four feature maps are fused by concatenation. And, a 1×1 convolution layer is applied to produce pointwise predictions in the parametric space. As a result, predicting a line becomes to predict a point that can be represented by (θ, r). The reason for predicting in the Hough space is that the borderlines indicate the boundary between the water surface and sewer wall and have more semantic information rather than structure information. The dashed line in Figure 7 is conducted only during testing to visualize the predictions.

4.2 | Spatial Net

The Spatial Net starts with two identical blocks to decrease the resolution. Each block has a convolution layer and a rectified linear unit (ReLU) nonlinear function. Then, another 11 residual blocks are used to increase the

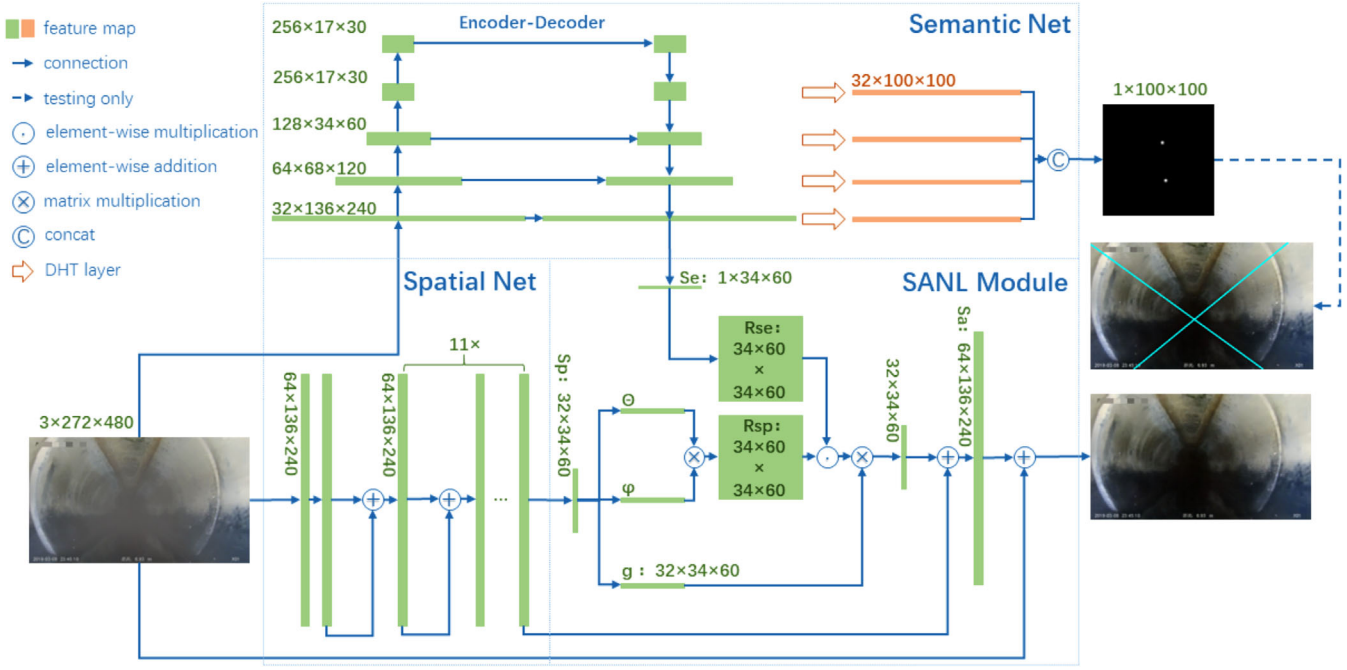


FIGURE 7 The overall architecture of the work contains two subnetworks and an SANL Module: (i) a Semantic Net, which predicts two borderlines; (ii) a Spatial Net, which is used to obtain the residual map; (iii) an SANL module, which realizes feature fusion between two subnetworks. The size format is channel × height × width. Note that the dashed line in the network indicates the step is only conducted during testing

TABLE 1 Convolution layers in the Semantic Net

Layer	Kernel	Stride	Dilation	Padding	Output
Input					3 × 272 × 480
Convolution1	4 × 4	2	1	1	32 × 136 × 240
Convolution2	4 × 4	2	1	1	64 × 68 × 120
Convolution3	4 × 4	2	1	1	128 × 34 × 60
Convolution4	4 × 4	2	1	1	256 × 17 × 30
Convolution5	3 × 3	1	2	2	256 × 17 × 30
Convolution6	3 × 3	1	4	4	256 × 17 × 30
Convolution7	3 × 3	1	2	2	256 × 17 × 30
Convolution8	4 × 4	2	1	1	128 × 34 × 60
Convolution9	4 × 4	2	1	1	64 × 68 × 120
Convolution10	4 × 4	2	1	1	32 × 136 × 240

receptive field without any downsampling operation on account of preserving spatial information. As for each residual block, there are a 3 × 3 dilated convolution (Chen, Papandreou, et al., 2018), an ReLU nonlinear function, another 3 × 3 dilated convolution, and a coordinate attention module (Hou et al., 2021). Then, a skip connection is adopted to add the input and output of each residual block. More details for convolution layers of 13 blocks in the Spatial Net are shown in Table 2.

Additionally, the role of the coordinate module in each residual block is to enhance positional information by embedding positional information into channel information. First, this module adopts two global pooling layers to the input with pooling kernels $(H, 1)$ and $(1, W)$ to yield two direction-aware tensors. Those two tensors are concatenated and sent to a 1 × 1 convolution followed by a nonlinear activation function to generate the intermediate feature map. Then, the intermediate feature map is split into two new tensors of the same size as the two direction-aware tensors. Finally, another 1 × 1 convolution and an activation layer are used to transform two new tensors into two attention weights.

TABLE 2 Convolution layers in the Spatial Net

Layer	Kernel	Stride	Dilation	Padding	Output
Input					3 × 272 × 480
Convolution1	4 × 4	2	-	1	32 × 136 × 240
Convolution2	1 × 1	1	-	0	64 × 164 × 240
Convolution3.1	3 × 3	1	1	1	64 × 136 × 240
Convolution3.2	3 × 3	1	1	1	64 × 136 × 240
Convolution4.1	3 × 3	1	1	1	64 × 136 × 240
Convolution4.2	3 × 3	1	1	1	64 × 136 × 240
Convolution5.1	3 × 3	1	2	2	64 × 136 × 240
Convolution5.2	3 × 3	1	2	2	64 × 136 × 240
Convolution6.1	3 × 3	1	2	2	64 × 136 × 240
Convolution6.2	3 × 3	1	2	2	64 × 136 × 240
Convolution7.1	3 × 3	1	4	4	64 × 136 × 240
Convolution7.2	3 × 3	1	4	4	64 × 136 × 240
Convolution8.1	3 × 3	1	8	8	64 × 136 × 240
Convolution8.2	3 × 3	1	8	8	64 × 136 × 240
Convolution9.1	3 × 3	1	4	4	64 × 136 × 240
Convolution9.2	3 × 3	1	4	4	64 × 136 × 240
Convolution10.1	3 × 3	1	2	2	64 × 136 × 240
Convolution10.2	3 × 3	1	2	2	64 × 136 × 240
Convolution11.1	3 × 3	1	2	2	64 × 136 × 240
Convolution11.2	3 × 3	1	2	2	64 × 136 × 240
Convolution12.1	3 × 3	1	1	1	64 × 136 × 240
Convolution12.2	3 × 3	1	1	1	64 × 136 × 240
Convolution13.1	3 × 3	1	1	1	64 × 136 × 240
Convolution13.2	3 × 3	1	1	1	64 × 136 × 240

The output is the multiplication of the weights and the input.

4.3 | SANL module

The SANL module aims to correlate the contextualized information and positional information over the entire feature maps. Before joining the SANL module, the Semantic Net adopts a residual block, a 3 × 3 convolution with padding = 1, a SELU, a 1 × 1 convolution with padding = 1 followed by a sigmoid function, to downsample the final output of the encoder–decoder network. The residual block is the same as the 11th block in the Spatial Net. The result is in the same size as the initial input image. To reduce the computation and memory cost, the height and width of the result are then both reduced as 1/8 of original values to get the feature map Se, which is realized by a 4 × 4 convolution with stride = 4. As for the output of the Spatial Net, an interpolation is applied to downsample it to get the feature map Sp.

Three 1 × 1 convolution layers are applied to the feature map Sp to separate the channels and get three components (θ, ϕ, g):

$$\theta = F_{\theta}(Sp) \tag{12}$$

$$\phi = F_{\phi}(Sp) \tag{13}$$

$$g = F_{g}(Sp) \tag{14}$$

where $F_{\theta/\phi/g}$ implies a convolution operation. In Figure 7, there are two relation maps at the same size of 34×60 × 34×60. The relation map Rsp is achieved by applying the matrix multiplication and Softmax normalization to θ and ϕ. This step can be formulated as follows:

$$Rsp = Softmax(\theta \otimes \phi) \tag{15}$$

Another relation map Rse is extracted from the feature map Se by computing the distance between each pair of pixels and conducting a Softmax operation. Considering huge spatial ranges of distance values, Rse is obtained in

exponential value:

$$Rse = Softmax(e^{-0.1|Se_i - Se_j|}) \tag{16}$$

where i ∈ {1, … , 34 × 60} and j ∈ {1, … , 34 × 60}.

Then the relation map Rse and Rsp are multiplied and normalized by a Softmax layer. This operation aims to model the feature interdependency. The result is multiplied with the component g then operated upon by another 1 × 1 convolution followed by the addition with Sp. The above steps can be formulated as:

$$Sa = F(Softmax(Rse \oplus Rsp) \otimes g) + Sp \tag{17}$$

where Sa indicates the feature map at the size of 64 × 136 × 120, F indicates a convolution layer. Ultimately, a 4 × 4 deconvolution layer, an RELU layer, and a 3 × 3 convolution layer are applied to resize Sa and the result is added to the initial input image to get the final output image.

4.4 | Loss functions

There are two supervision signals in SANL-Net. The structure loss of the Semantic Net is between the ground-truth map H and the prediction $\widetilde{H}$ in the Hough space. An operation is conducted to smooth the ground-truth map H:

$$H^K = H \circledast K \tag{18}$$

where K is a 5 × 5 Gaussian kernel, ⊛ denotes the convolution operation. Then, the structure loss $\mathcal{L}_S$ is given by a cross-entropy loss:

$$\mathcal{L}_S = -\sum_i H_i^K \cdot log(\widetilde{H}_i) + (1 - H_i^K) \cdot log(1 - \widetilde{H}_i) \tag{19}$$

where i indicates the class index. The dehazing loss $\mathcal{L}_D$ of the SANL module is between the clean ground-truth I and the dehazed image $\widetilde{I}$, which are all already normalized to [0,1]. The dehazing loss $\mathcal{L}_D$ can be given by:

$$\mathcal{L}_D = \sum_{c \in (R,G,B)} \sum_x |I(x)_c - \widetilde{I}(x)_c| \tag{20}$$

where c indicates the channel in the RGB space and x indicates the pixel position in the image. The total loss $\mathcal{L}$ can be given by:

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_D \tag{21}$$

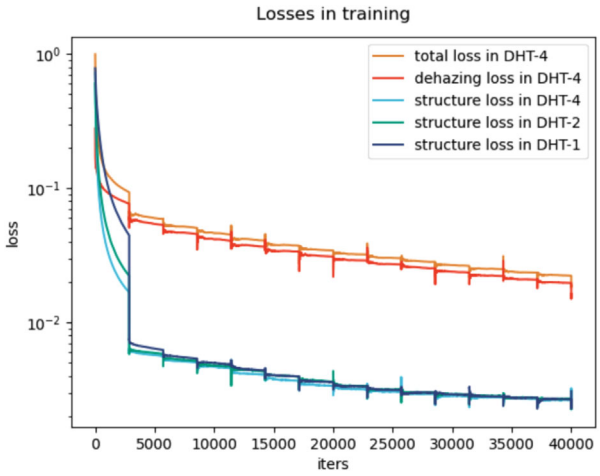


FIGURE 8 Training losses. Note that DHT-4 indicates SANL-Net while DHT-2 and DHT-1 are for ablation studies

5 | EXPERIMENTS

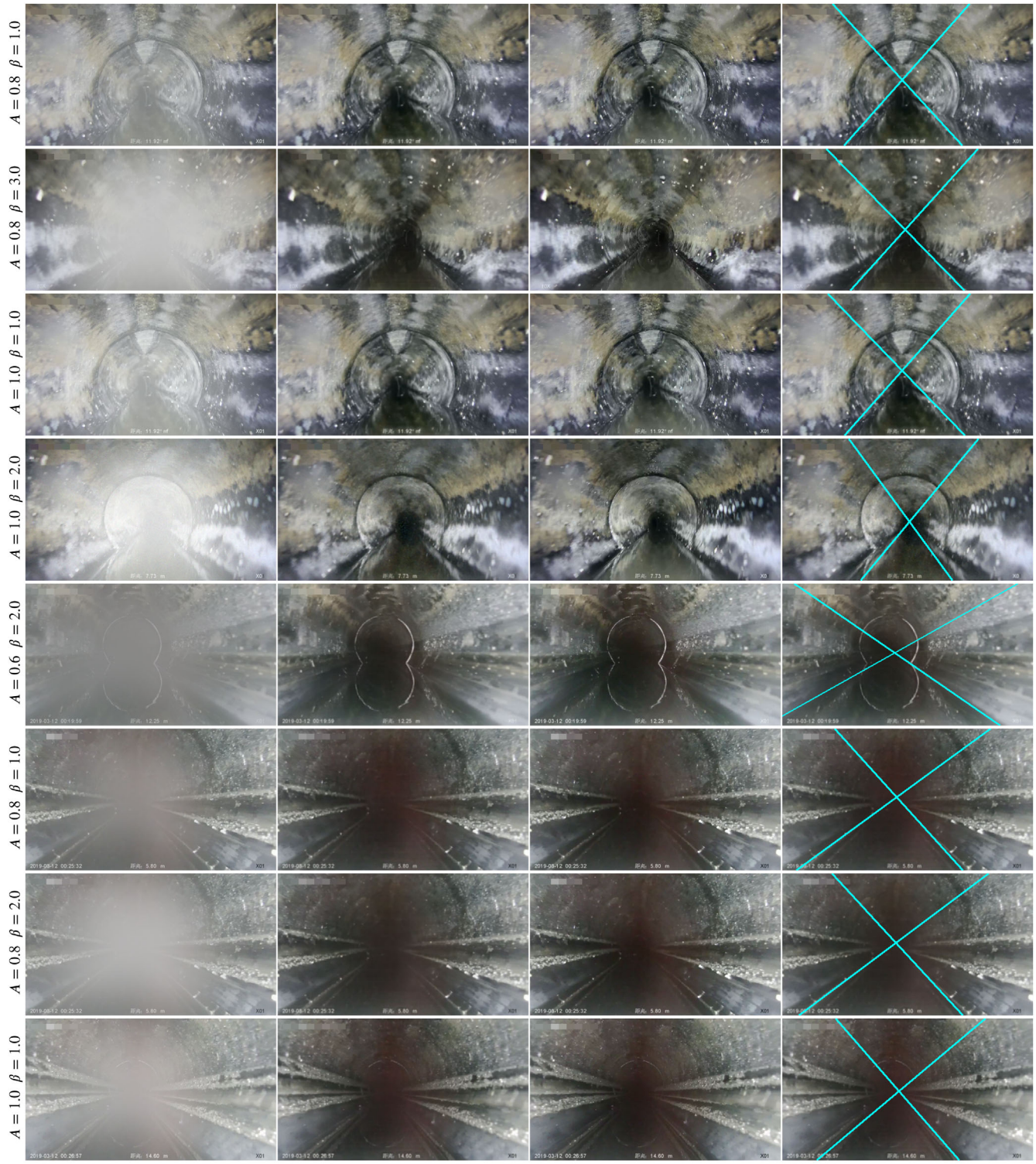
5.1 | Training strategies

The whole network is deployed in the PyTorch framework (Paszke et al., 2019) while DHT layers are implemented in native CUDA programming and others are implemented based on framework level Python API. Adam optimizer is used with decay and momentum of 0.9 at the training stage. The initial learning rate is set to 5e-4 and is decayed with the "poly" policy with the power of 0.9. The batch size is 4 and the iteration step is 40,000. A single NVidia 3090Ti GPU is used for training and testing. The convergence situation of $\mathcal{L}$, $\mathcal{L}_S$, $\mathcal{L}_D$ in training can be seen in Figure 8. All losses are decreased greatly within the first 2500 iterations.

5.2 | Verification

5.2.1 | Verification in the testing set

SANL-Net generates two outputs at the verification stage: two borderlines and a dehazed image. Figure 9 shows some verification examples of the testing set. The higher atmosphere light A and scattering coefficient β result in a hazier image (see the sixth and seventh rows). But the dehazing result does not depend on the degradation degree of inputs. For example, the dehazed image in the sixth row is not clearer than the one in the seventh row. Because peak signal to noise ratio (PSNR) of the six row is 31.02 while it is 31.17 in the seventh row. Actually, the more accurate the borderlines prediction is, the better the dehazing result is (see the first four and last four rows). As



(a) hazy image (b) dehazed image (c) ground-truth (d) borderlines

FIGURE 9 Results for the testing set

the S2B dataset contains synthesized hazy inputs with relatively accurate depth maps, the models can handle the thick haze around the vanishing point. Moreover, the fourth row shows the model's robustness in the extremely hazy images.

Table 3 shows the comparison between SANL-Net and other state-of-the-art dehazing methods. The metrics include mean square error (MSE), PSNR, structural similarity index measure (SSIM) (Sara et al., 2019; Wang et al., 2004), and parameters of different models. GCANet is the

TABLE 3 Methods comparison. MSE is calculated by the squared intensity differences of hazy and dehazed image pixels. PSNR is the ratio between the maximal of the image data and MSE. SSIM gives normalized mean value of structural similarity between the two images bold numbers: the first two results)

Method	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
SANL-Net	**147**	**27.28**	**0.8963**	15.47M

same as the original sewer image dehazing work (Li et al., 2022). For a fair comparison (Luo et al., 2017; Xiao et al., 2021), other models are trained with the same epochs (14 epochs) as SANL-Net. And, the parameters of SANL-Net do not include parameters of DHT layers as these layers only play a part at the training stage. The work outperforms other works in most cases and has a relatively lower parameter size.

5.2.2 | Verification in the wild data

The wild data come from the real hazy photos without ground truth. Figure 9 shows some verification examples. The centers of most real photos are full of thick haze, which is consistent with the synthesized images. So although trained with synthesized images, SANL-Net is found to perform well on real photos. Rich and saturated colors have remained and color tones are also relatively realistic. And dehazing effect is also related to borderline prediction. For example, an accurate prediction in the second row corresponds to a clear dehazed image.

5.3 | Application in the Pipe dataset

Haze affects the visibility of CCTV images and declines the performance of sewer inspection. To assess how the proposed method contributes to improving the accuracy of sewer inspection, different defects detection methods (i.e., semantic segmentation, object localization, and image classification) are applied to detect defects in synthesized hazy images, in dehazed counterparts produced from SANL-Net, and in ground-truth counterparts for comparison. First, SANL-Net is trained on the S2B dataset. Second, detection models are trained on the training sets of sewer defect datasets including clean images. Finally, the trained models are used to detect defects in the validation set

TABLE 4 Sewer defect classes

Dataset	Code	Description
Pipe	JO	Joint offset
	IL	Intruding lateral
	IN	Infiltration

of the sewer defect datasets, including synthesized hazy images, dehazed images by SANL-Net, and clean ground-truth images. Synthesized images are obtained by the same method used to produce the S2B data set.

These applications aim to analyze how haze affects different tasks. And the application results are assessed by three quantities: the hazy value HV, dehazed value DV, and nondehazed value UV, which can be given by:

$$HV = clean - hazy$$

$$DV = dehazed - hazy$$

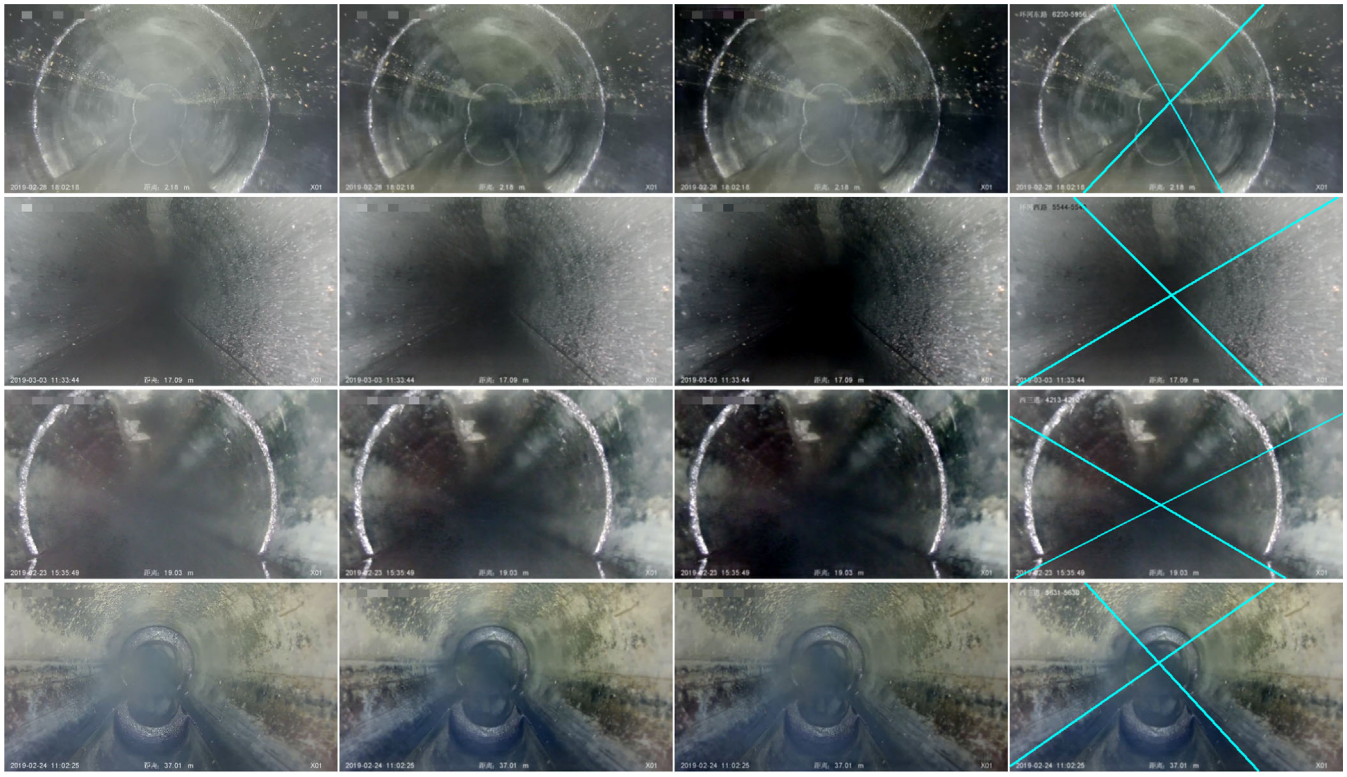
$$UV = clean - dehazed$$

For example, "clean" indicates the original detection metric (such as mean intersection over union [mIoU]%, mean average precision [mAP]%, and F1%) for clean images. A higher HV indicates that haze affects the detection performance heavily. A higher DV indicates a better dehazing performance while a higher UV indicates a worse dehazing performance. These three values are expected to be positive.

This section illustrates the application results of the dataset in our previous work (Pan et al., 2020) named Pipe dataset. Table 4 shows the code description for every defect class in the Pipe dataset. Because of their different source, the aspect ratios of images in the S2B and Pipe differ. So two strategies are used during depth estimation for the Pipe dataset: one is estimating the depth map at the resolution of 256×256 as same as the Pipe dataset; the other is at the resolution of 480×272 as same as the training dataset. Figure 11 shows these two strategies.

5.3.1 | Semantic segmentation

Table 5 offers semantic segmentation results of the Pipe data set in mIoU% by different semantic segmentation models. The caption "w/o resizing" means the depth map is estimated using strategy shown in Figure 11a while "w/ resizing" corresponds to Figure 11b. In the original works where they appear, PipeUNet (Pan et al., 2020) is for sewer defect segmentation, FCN (Yang et al., 2018) is for crack inspection, and DeepLabv3+ (Chen, Zhu, et al., 2018) is for general segmentation. In Table 5, although depth maps of images w/o resizing are estimated at a resolution



(a) hazy image (b) SANL-Net (c) SANL-Net w/o d (d) SANL-Net (borderlines prediction)

FIGURE 10 Results for the wild data. "SANL-Net w/o d" means the dataset used to train is synthesized by adding even haze without depth information and more details can be seen in Section 5.6. "SANL-Net (borderlines prediction)" represents the prediction result of two water borderlines

(a) w/o resizing (b) w/ resizing

FIGURE 11 Two strategies applied during depth estimation for the Pipe data set: (a) is at the resolution of 256 × 256, (b) is at the resolution of 480 × 272

different from the S2B dataset, SANL-Net still performs well. mIoU% of images w/o resizing is always higher than image w/ resizing, which implies resizing operation results in deformation and information loss. What is unexpected is that mIoU% of clean images w/o resizing or w/ resizing is different. It is probably that semantic segmentation models have better performances on the original resolution of images.

DeepLabv3+ always performs better than the other two models in mean mIoU% of three image types (hazy, dehazed, and clean) and mean *DV*. FCN has a lower mean *HV*, which implies FCN shows great robustness to haze. And defect class JO is easier to be recognized than other defect classes. In terms of *DV* for different classes, PipeUNet and DeepLabv3+ perform better in class IL while FCN performs better in class IN. What's more, class IL always has a higher *HV* than other defect classes, which means that mIoU% of this class is affected significantly by the unrealistic color zones caused by haze.

5.3.2 | Object localization

Table 6 shows object localization results of the Pipe data set in mAP% by different object localization models. In their originating works, YOLOv5s[1] is for general localization while faster R-CNN (Kumar et al., 2020) and SSD (Kumar et al., 2020) are for sewer defect localization. In Table 6, at the same resolution as same as the S2B data set, images w/ resizing achieve higher *DV* again in object localization, which is similar to semantic segmentation models. But objection localization models show better adaptive capabilities to the resolution change than semantic segmentation models.

[1] The code is from https://github.com/ultralytics/yolov5

TABLE 5 Semantic segmentation results of the Pipe data set in mIoU%

		PipeUNet			FCN			DeepLabv3+		
	Class	Hazy^{HV}	Dehazed^{DV}	Clean^{UV}	Hazy^{HV}	Dehazed^{DV}	Clean^{UV}	Hazy^{HV}	Dehazed^{DV}	Clean^{UV}
w/o resizing	JO	$47.83^{14.42}$	$57.43^{9.6}$	$62.25^{4.82}$	$57.89^{5.47}$	$59.78^{1.89}$	$63.36^{3.58}$	$60.58^{9.38}$	$63.52^{2.94}$	$69.96^{6.44}$
	IL	$21.68^{26.45}$	$32.77^{11.09}$	$48.13^{15.36}$	$24.56^{21.45}$	$36.26^{11.7}$	$46.01^{9.75}$	$16.54^{28.75}$	$35.66^{19.12}$	$45.29^{9.63}$
	IN	$8.77^{20.02}$	$18.4^{9.63}$	$28.79^{10.39}$	$30.53^{20.41}$	44.53^{14}	$50.94^{6.41}$	$42.12^{16.71}$	$53.45^{11.33}$	$58.83^{5.38}$
	Mean[a]	$26.09^{20.3}$	$36.2^{10.11}$	$46.39^{10.19}$	$37.66^{15.78}$	$46.86^{9.2}$	$53.44^{6.58}$	$39.75^{18.28}$	$50.87^{11.13}$	$58.03^{7.16}$
w/ resizing	JO	$29.08^{29.8}$	$52.13^{23.05}$	$58.88^{6.75}$	$47.85^{12.15}$	$52.89^{5.04}$	$60^{7.11}$	$40.24^{25.6}$	$56.15^{15.91}$	$65.84^{9.69}$
	IL	$10.67^{30.37}$	$34.94^{24.27}$	$41.04^{6.1}$	$16.03^{21.11}$	$33.92^{17.89}$	$37.14^{3.22}$	$9.49^{35.11}$	$40.71^{31.22}$	$44.6^{3.89}$
	IN	$5.53^{23.32}$	$18.43^{12.9}$	$28.85^{10.42}$	$17.77^{34.02}$	$48.86^{31.09}$	$51.79^{2.93}$	$34.58^{27.59}$	$56.94^{22.36}$	$62.17^{5.23}$
	Mean[a]	$15.09^{27.83}$	$35.17^{20.07}$	$42.92^{7.75}$	$27.22^{22.42}$	$45.22^{18.01}$	$49.64^{4.42}$	$28.1^{29.43}$	$51.27^{23.16}$	$57.53^{6.26}$

[a]Mean value for all defect classes.

TABLE 6 Object localization results of the Pipe dataset in mAP%

		YOLOv5s			Faster R-CNN			SSD		
	Class	Hazy^{HV}	Dehazed^{DV}	Clean^{UV}	Hazy^{HV}	Dehazed^{DV}	Clean^{UV}	Hazy^{HV}	Dehazed^{DV}	Clean^{UV}
w/o resizing	JO	$74.16^{5.88}$	$79.03^{4.87}$	$80.04^{1.01}$	$62.65^{9.49}$	$64.74^{2.09}$	$72.14^{7.4}$	$61.99^{8.11}$	$68.65^{6.66}$	$70.1^{1.45}$
	IL	$42.74^{20.05}$	$59.96^{17.22}$	$62.79^{2.83}$	$28.8^{38.72}$	$59.61^{30.81}$	$67.52^{7.91}$	$31.44^{22.79}$	$52.13^{20.69}$	$54.23^{2.1}$
	IN	$43.88^{19.06}$	$59.23^{15.35}$	$62.94^{3.71}$	$33.16^{18.28}$	$41.14^{7.98}$	$51.44^{10.3}$	$28.28^{23.23}$	$44.77^{16.49}$	$51.51^{6.74}$
	Mean[a]	53.59^{15}	$66.07^{12.48}$	$68.59^{2.52}$	$41.54^{22.16}$	$55.16^{13.63}$	$63.7^{8.54}$	$40.57^{18.04}$	$55.18^{14.61}$	$58.61^{3.43}$
w/ resizing	JO	$61.49^{15.28}$	$74.63^{13.14}$	$76.77^{2.14}$	$42.19^{4.16}$	$43.37^{1.18}$	$46.35^{2.98}$	$57.03^{13.32}$	$68.14^{11.11}$	$70.35^{2.21}$
	IL	$44.48^{18.77}$	$60.76^{16.28}$	$63.25^{2.49}$	$29.52^{34.17}$	$56.36^{26.84}$	$63.69^{7.33}$	$29.55^{25.03}$	$53.77^{24.22}$	$54.58^{0.81}$
	IN	$38.24^{34.22}$	$63.84^{25.6}$	$72.46^{8.62}$	$21.44^{25.6}$	$42.25^{20.81}$	$47.04^{4.79}$	$30.18^{21.01}$	$50.73^{20.55}$	$51.19^{0.46}$
	Mean[a]	$48.07^{22.76}$	$66.41^{18.34}$	$70.83^{4.42}$	$31.05^{21.31}$	$47.33^{16.28}$	$52.36^{5.03}$	$38.92^{19.79}$	$57.55^{18.63}$	$58.71^{1.16}$

[a]Mean value for all defect classes.

YOLOv5s performs better in mean mAP% of three image types. SSD performs better in mean *DV*, which implies that SANL-Net shows a significant enhancement effect on FCN. The higher mean *UV* of faster R-CNN also indicates SANL-Net cannot improve FCN performance effectively. And class JO is always easier to be localized than other classes in most situations. This situation also happens in semantic segmentation models. Therefore, improvement is needed in both semantic segmentation and object localization of class IL and IN.

5.4 | Application in sewer-ML dataset

The Sewer-MLdataset (Haurum & Moeslund, 2021) is used for image classification. Depth maps of synthesized hazy images are estimated at the resolution as same as the Sewer-ML dataset. This section illustrates image classification results of the Sewer-ML data set.

Table 7 shows classification results by different models. In their original works, GoogleNet InceptionV3 (Chen et al., 2018) and IDCNN&NDCNN (Xie et al., 2019) are for sewer defect classification while ResNet-101 (He et al., 2016) is for general classification. In this work, the positive mean *DV* of three classification models implies the effectiveness of SANL-Net. ResNet-101 outperforms the other two methods in terms of F1% of three image types. But the classification performance is significantly influenced by thick haze. So F1% of dehazed images produced by SANL-Net cannot be comparable with F1% of clean images. For example, there are also negative *HV*, *DV*, and *UV* for some defect classes, such as *DV* of class IN by GoogleNet InceptionV3, *HV* and *UV* of class PF the GoogleNet InceptionV3. One reason for this situation is that the total number of defect classes is up to 15, while images of some defect classes used in the work are not enough.

5.5 | Application in the wild data

The wild data represent real hazy photos without defect labels. So improvement by SANL-Net cannot be assessed in terms of the above three quantitative metrics anymore. Therefore, three trained semantic segmentation models and three trained object localization models are applied in real hazy photos and their dehazed counterparts produced by SANL-Net.

TABLE 7 Image classification results of the sewer-ML dataset in F1%

	GoogleNet InceptionV3			ResNet-101			IDCNN&NDCNN		
Class	**Hazy^{HV}**	**Dehazed^{DV}**	**Clean^{UV}**	**Hazy^{HV}**	**Dehazed^{DV}**	**Clean^{UV}**	**Hazy^{HV}**	**Dehazed^{DV}**	**Clean^{UV}**
RB	$21.87^{9.24}$	$26.2^{4.33}$	$31.11^{4.91}$	$23.1^{4.81}$	$18.45^{-4.65}$	$27.91^{9.46}$	$12.86^{9.57}$	$11.16^{-1.7}$	$22.43^{11.27}$
OB	$60.43^{9.29}$	$67.75^{7.32}$	$69.72^{1.97}$	$66.46^{8.77}$	$72.36^{5.9}$	$75.23^{2.87}$	$27.3^{34.82}$	$43.27^{15.97}$	$62.12^{18.85}$
PF	$30.47^{24.08}$	$48.54^{18.07}$	$54.55^{6.01}$	$40.54^{18.92}$	$54.74^{14.2}$	$59.46^{4.72}$	$10.34^{33.66}$	$19.58^{9.24}$	$44^{29.42}$
DE	$48.66^{18.01}$	$63.46^{14.8}$	$66.67^{3.21}$	$63.1^{17.71}$	$71.79^{8.69}$	$80.81^{9.02}$	$5.29^{29.28}$	$40^{34.71}$	$34.57^{-5.43}$
FS	$63.94^{12.25}$	$68.25^{4.31}$	$76.19^{7.94}$	$64.41^{10.78}$	$72.51^{8.1}$	$75.19^{2.68}$	$27.89^{39.64}$	$41.63^{13.74}$	$67.53^{25.9}$
IS	$7.78^{13.27}$	$21.43^{13.65}$	$21.05^{-0.38}$	$15.85^{5.58}$	$29.25^{13.4}$	$21.43^{-7.82}$	$1.18^{10.58}$	$8.1^{6.92}$	$11.76^{3.66}$
RO	$25.77^{11.27}$	$39.11^{13.34}$	$37.04^{-2.07}$	$40.22^{-0.22}$	$48^{7.8}$	40^{-8}	$10.21^{6.81}$	$20.45^{10.24}$	$17.02^{-3.43}$
IN	$39.01^{-3.45}$	$38.06^{-0.95}$	$35.56^{-2.5}$	$42.49^{-0.63}$	$45.27^{2.78}$	$41.86^{-3.41}$	$19.83^{-1.81}$	$13.02^{-6.81}$	18.02^{5}
AF	$17.22^{16.11}$	$11.43^{-5.79}$	$33.33^{21.9}$	$18.95^{13.48}$	$9.34^{-9.61}$	$32.43^{23.09}$	$6.15^{2.01}$	$9.84^{3.69}$	$8.16^{-1.68}$
BE	$43.48^{6.52}$	$41.39^{-2.09}$	$50^{8.61}$	$55.99^{3.53}$	$47.78^{-8.21}$	$59.52^{11.74}$	$28.16^{7.74}$	$34.81^{6.65}$	$35.9^{1.09}$
FO	$11.06^{6.33}$	$17.57^{6.51}$	$17.39^{-0.18}$	$7.37^{13.68}$	$19.9^{12.53}$	$21.05^{1.15}$	$14.41^{-8.01}$	$7.63^{-6.78}$	$6.4^{-1.23}$
GR	$28.9^{19.38}$	$52.02^{23.12}$	$48.28^{-3.74}$	$42.58^{7.42}$	$53.82^{11.24}$	$50^{-3.82}$	$11.16^{4.37}$	$17.41^{6.25}$	$15.53^{-1.88}$
PH	$37.08^{9.07}$	$62.43^{25.35}$	$46.15^{-16.28}$	$42.65^{1.79}$	$57.75^{15.1}$	$44.44^{-13.31}$	$4.49^{9.8}$	$12.59^{8.1}$	$14.29^{1.7}$
OP	$19.57^{16.79}$	$23.29^{3.72}$	$36.36^{13.07}$	$44.83^{-8.47}$	$28.8^{-16.03}$	$36.36^{7.56}$	$5.26^{-0.13}$	$0.79^{-4.47}$	$5.13^{4.34}$
OK	$22.22^{-7.93}$	$14.4^{-7.82}$	$14.29^{-0.11}$	$65.93^{-5.93}$	$49.57^{-16.36}$	$60^{10.43}$	$2.19^{5.29}$	$3.66^{1.47}$	$7.48^{3.82}$
Mean[a]	$31.83^{10.68}$	$39.69^{7.86}$	$42.51^{2.82}$	$42.3^{6.08}$	$45.29^{2.99}$	$48.38^{3.09}$	$12.45^{12.24}$	$18.93^{6.48}$	$24.69^{5.76}$

[a]Mean value for all defect classes.

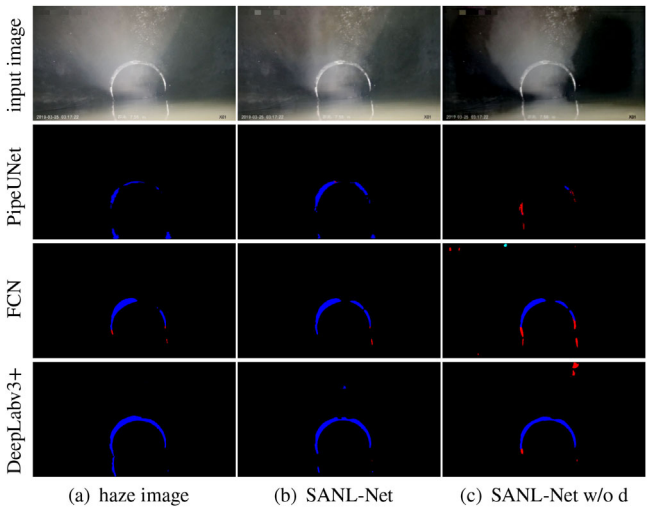


FIGURE 12 Semantic segmentation results of the wild data

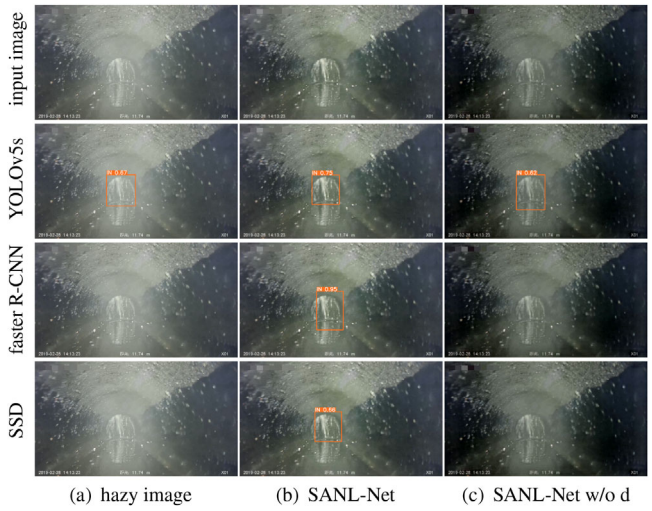


FIGURE 13 Object localization results of the wild data

Figure 12 shows semantic segmentation results of the wild data, in which red and blue pixels, respectively, indicate class IN and JO. DeepLabv3+ shows excellent performance in every image type (hazy and dehazed). Class JO of dehazed images can be recognized more accurately by every model. Although PipeUNet produces poor results, its improvement from the hazy image to the dehazed one is the greatest.

Figure 13 shows object localization results. YOLOv5s shows its superiority in every image type (hazy and dehazed). Class JO of dehazed images achieves higher confidence or can be localized.

5.6 | Ablation studies

Ablation studies are applied to verify the effectiveness of depth estimation and components of network architecture. The metrics MSE, PSNR, and SSIM are used to assess the quantitative effects.

5.6.1 | W/ or w/o depth estimation

To prove the effectiveness of depth estimation, haze is added to the clean image without depth information.

TABLE 8 Ablation studies. Recall% is about the predictions of two borderlines.

coor	DHT-1	DHT-2	DHT-4	MSE	PSNR	SSIM	Recall%
				159	26.95	0.8921	
✓				156	27.01	0.8956	
✓	✓			177	26.59	0.8921	66.32
✓		✓		155	27.13	0.8953	66.93
✓			✓	147	27.28	0.8963	67.75

Those synthesized images and clean ones construct a new dehazing dataset. The network in Figure 7 trained by this dataset is called SANL-Net w/o d.

Figure 10 shows the dehazing results of SANL-Net w/o d. The lack of depth information not only makes SANL-Net w/o d unable to balance thick haze in the image center (see the first row and the fourth row) but also causes unreal color zones at the border of the image (the first three rows).

Figure 12 and Figure 13 show application results of SANL-Net w/o d. In Figure 12, SANL-Net w/o d darkens some regions and creates fake color tones, which makes the segmentation result of the dehazed image even worse than the original image's. In Figure 13, the localization results of the dehazed image have not been improved at all. And, YOLOv5s even obtain a declined performance in the dehazed image.

5.6.2 | Network architecture

In Table 8, "coor" represents the coordinate attention module, which proved to be useful for image dehazing. DHT-n indicates the last n feature maps in the decoder network are used to produce the final prediction map in Hough space. Figure 8 shows that more feature maps are better for the convergence speed of the structure loss. Table 8 shows more feature maps could improve the recall% of borderlines. It is interesting to find that DHT-1 and DHT-2 are not enough to guide the Dehazing Net and their participation misleads the model to yield a low SSIM value.

6 | CONCLUSION

To solve the sewer image dehazing problem caused by complex sewer internal environment, this work first proposes a depth estimation method using monocular cues in the cylindrical sewer. A dehazeing dataset called S2B dataset is built by synthesizing hazy images with clean images and their depth maps. The proposed network called SANL-Net acts as a joint network between image dehazing and water borderlines detection and is trained with the S2B dataset. Compared with several state-of-the-art methods, SANL-Net shows its superiority for sewer images with 147 in MSE, 27.28 in PSNR, 0.8963 in SSIM, and 15.47M in parameters. In addition, the outstanding performance on real photo dehazing indicates the accuracy of depth estimation. Application experiments show how SANL-Net contributes to different tasks for sewer inspection, such as an increase of 23.16% in mIoU for semantic segmentation. Ablation studies show the indispensability of different components of this method. Therefore, the proposed method can be used as a preprocessing step of sewer automatic defects detection tools. In future, more powerful machine learning methods, such as probabilistic neural network and dynamic neural network, will be considered in sewer image dehazing to realize a real-time and more accurate sewer inspection.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of Tianjin (No. 21JCYBJC00640), the Key Research and Development Plan Support Program of Tianjin (No. 20YFZCSN01080), the Key Research and Development Program in Anhui Province (No. 202104i07020012), and National Cultural and Tourism Science and Technology Innovation Project of China (No. 2021-97) University-Industry Collaborative Education Program (No. 202102012011).

REFERENCES

Angot, L. J. J., Huang, W. J., & Liu, K. C. (2010). A 2D to 3D video and image conversion technique based on a bilateral filter. In Baskurt, A. M. (Ed.), *Three-dimensional image processing (3DIP) and applications* (Vol. 7526, pp. 88–97). International Society for Optics and Photonics, SPIE.

Anvari, Z., & Athitsos, V. (2020). Dehaze-GLCGAN: unpaired single image de-hazing via adversarial training. arXiv preprint arXiv:2008.06632.

Berman, D., treibitz, T., & Avidan, S. (2016). Non-local image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 1674-1682)*. IEEE.

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable con- volution for semantic image segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Com- puter vision – eccv 2018 (pp. 833–851)*. Cham: Springer International Publishing.

Chen, D., He, M., Fan, Q., Liao, J., Zhang, L., Hou, D., Yuan, L., & Hua, G. (2019). Gated context aggregation network for image dehazing and deraining. In *2019 IEEE winter conference on applications of computer vision (WACV)* (pp. 1375–1383). IEEE.

Chen, K., Hu, H., Chen, C., Chen, L., & He, C. (2018). An intelligent sewer defect detection method based on convolutional neural network. In *2018 IEEE International conference on information and automation (ICIA)* (pp. 1301–1306). IEEE.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Cheng, J. C., & Wang, M. (2018). Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques. *Automation in Construction*, *95*, 155–171.

Cooper, D., Pridmore, T., & Taylor, N. (1998). Towards the recovery of extrinsic camera parameters from video records of sewer surveys. *Machine Vision and Applications*, *11*(2), 53–63.

Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., & Yang, M. H. (2020). Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) (pp. 2157-2167)*. IEEE.

Ge, G., Cheng, Z., Ke, P., & Yu, J. (2017). Depth map extracting based on geometric perspective: An applicable 2d to 3d conversion technology. In *2017 10th international Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* (pp. 1–5).

Guo, W., Soibelman, L., & Garrett, J. H. (2008). Imagery enhancement and interpretation for remote visual inspection of aging civil infrastructure. *Tsinghua Science & Technology*, *13*, 375–380.

Guo, W., Soibelman, L., & Garrett, J. H. (2009). Automated defect detection for sewer pipeline inspection and condition assessment. *Automation in Construction*, *18*(5), 587–596.

Hassanpour, A., Moradikia, M., Adeli, H., Khayami, S. R., & Shamsinejadbabaki, P. (2019). A novel end-to-end deep learning scheme for classifying multi-class motor imagery electroencephalography signals. *Expert Systems*, *36*(6), e12494.

Haurum, J. B., & Moeslund, T. B. (2021). Sewer-ML: A multi-label sewer defect classification dataset and benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 13456–13467). IEEE.

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision (ICCV) (pp. 2961-2969)*. IEEE.

He, K., Sun, J., & Tang, X. (2011). Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(12), 2341–2353.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 770-778)*. IEEE.

Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 13713–13722). IEEE.

Hu, X., Zhu, L., Wang, T., Fu, C. W., & Heng, P. A. (2021). Single-image real-time rain removal based on depth-guided non-local features. *IEEE Transactions on Image Processing*, *30*, 1759–1770.

Iyer, S., & Sinha, S. K. (2006). Segmentation of pipe images for crack detection in buried sewers. *Computer-Aided Civil and Infrastructure Engineering*, *21*(6), 395–410.

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-normalizing neural networks. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 972–981).

Kumar, S. S., Wang, M., Abraham, D. M., Jahanshahi, M. R., Iseley, T., & Cheng, J. C. P. (2020). Deep learning-based automated detection of sewer defects in CCTV videos. *Journal of Computing in Civil Engineering*, *34*(1), 04019047.

Kunzel, J., Werner, T., Eisert, P., & Waschnewski, J. (2018). Automatic analysis of sewer pipes based on unrolled monocular fisheye images. In *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 2019–2027). IEEE.

Li, B., Peng, X., Wang, Z., Xu, J., & Feng, D. (2017). AOD-Net: All-in-one dehazing network. In *Proceedings of the IEEE international conference on computer vision* (pp. 4770–4778). IEEE.

Li, C., He, T., Wang, Y., Zhang, L., Liu, R., & Zheng, J. (2020). Pipeline image haze removal system using dark channel prior on cloud processing platform. *International Journal of Computational Science and Engineering*, *22*(1), 84–95.

Li, Y., Wang, H., Dang, L. M., Piran, M. J., & Moon, H. (2022). A robust instance segmentation framework for underground sewer defect detection. *Measurement*, 190, 110727.

Luo, J. H., Wu, J., & Lin, W. (2017). Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE international conference on computer vision* (pp. 5058–5066). IEEE.

Mao, J., Xiao, T., Jiang, Y., & Cao, Z. (2017). What can help pedestrian detection? In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 3127-3136)*. IEEE.

Martins, G. B., Papa, J. P., & Adeli, H. (2020). Deep learning techniques for recommender systems based on collaborative filtering. *Expert Systems*, *37*(6), e12647.

McCartney, E. J. (1976). *Optics of the atmosphere: Scattering by molecules and particles*. New York.

Narasimhan, S. G., & Nayar, S. K. (2001). Removing weather effects from monochrome images. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001* (Vol. 2, pp. II–II). IEEE.

Narasimhan, S. G., & Nayar, S. K. (2002). Vision and the atmosphere. *International Journal of Computer Vision*, *48*(3), 233–254.

Narasimhan, S. G., & Nayar, S. K. (2003). Contrast restoration of weather degraded images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(6), 713–724.

Nayar, S. K., & Narasimhan, S. G. (1999). Vision in bad weather. In *Proceedings of the seventh IEEE international conference on computer vision* (Vol. 2, pp. 820–827). IEEE.

Pan, G., Zheng, Y., Guo, S., & Lv, Y. (2020). Automatic sewer pipe defect semantic segmentation based on improved u-net. *Automation in Construction*, *119*, 103383.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Pinheiro, P. O., Lin, T. Y., Collobert, R., & Dollár, P. (2016). Learning to refine object segments. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision—ECCV 2016* (pp. 75–91). Springer International Publishing.

Rafiei, M., & Adeli, H. (2018). Novel machine learning model for construction cost estimation taking into account economic variables and indices. *Journal of Construction Engineering and Management*, *144*(12), 04018106.

Sara, U., Akter, M., & Uddin, M. S. (2019). Image quality assessment through FSIM, SSIM, MSE and PSNR—A comparative study. *Journal of Computer and Communications*, *7*(3), 8–18.

Su, J., Chen, C., Zhang, K., Luo, J., Wei, X., & Wei, X. (2021). Structure guided lane detection. In Z. H. Zhou (Ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21* (pp. 997–1003). International Joint Conferences on Artificial Intelligence Organization. Main Track.

Wang, M., & Cheng, J. C. (2020). A unified convolutional neural network integrated with conditional random field for pipe defect segmentation. *Computer-Aided Civil and Infrastructure Engineering*, *35*(2), 162–177.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wu, Y., & He, K. (2018). Group normalization. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Com- puter vision – eccv 2018 (pp. 3–19)*. Cham: Springer International Publishing.

Xiao, T., Dollar, P., Singh, M., Mintun, E., Darrell, T., & Girshick, R. (2021). Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, *34*, 30392–30400.

Xie, Q., Li, D., Xu, J., Yu, Z., & Wang, J. (2019). Automatic detection and classification of sewer defects via hierarchical deep learning. *IEEE Transactions on Automation Science and Engineering*, *16*(4), 1836–1847.

Yang, X., Li, H., Yu, Y., Luo, X., Huang, T., & Yang, X. (2018). Automatic pixel-level crack detection and measurement using fully convolutional network. *Computer-Aided Civil and Infrastructure Engineering*, *33*(12), 1090–1109.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M. H., & Shao, L. (2021). Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 14821–14831). IEEE.

Zhang, J., Xu, Y., Ni, B., & Duan, Z. (2018). Geometric constrained joint lane segmentation and lane boundary detection. In *Proceedings of the European conference on computer vision (ECCV) (pp. 486-512)*.

Zhang, Z., Qiao, S., Xie, C., Shen, W., Wang, B., & Yuille, A. L. (2018). Single-shot object detection with enriched semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 5813-5821)*. IEEE.

Zhao, K., Han, Q., Zhang, C. B., Xu, J., & Cheng, M. M. (2021). Deep Hough transform for semantic line detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Zhu, Q., Mai, J., & Shao, L. (2015). A fast single image haze removal algorithm using color attenuation prior. *IEEE Transactions on Image Processing*, *24*(11), 3522–3533.

How to cite this article: Xia, Z., Guo, S., Sun, D., Lv, Y., Li, H., & Pan, G. (2023). Structure-aware dehazing of sewer inspection images based on monocular depth cues. *Computer-Aided Civil and Infrastructure Engineering*, *38,* 762–778. https://doi.org/10.1111/mice.12900

APPENDIX A: MATHEMATICAL DERIVATION FOR DEPTH ESTIMATION

In Figure 2, specify that

- $A^w = [x, y, z]^T$
- $B^w = [-x, y, z]^T$
- $O^w = [0, 0, z]^T$
- $V^w = [0, t, z]^T$

Because A and B are in the sewer wall, it is easy to get:

$$x^2 + y^2 = \mathcal{R}^2 \tag{A1}$$

After transformations of the camera matrix, the pixel coordinates of these points are as follows:

$$A^{im} = \frac{1}{z}\begin{bmatrix} fxcos\theta - fysin\theta + ftsin\theta \\ fxsin\theta + fycos\theta - ftcos\theta \end{bmatrix} \tag{A2}$$

$$B^{im} = \frac{1}{z}\begin{bmatrix} -fxcos\theta - fysin\theta + ftsin\theta \\ -fxsin\theta + fycos\theta - ftcos\theta \end{bmatrix} \tag{A3}$$

$$O^{im} = \frac{1}{z}\begin{bmatrix} ftsin\theta \\ -ftcos\theta \end{bmatrix} \tag{A4}$$

$$V^{im} = \frac{1}{z}\begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{A5}$$

The euclidean distances between these points can be given by:

$$O^{im}A^{im} = \frac{1}{z}\sqrt{f^2\mathcal{R}^2} \tag{A6}$$

$$O^{im}B^{im} = \frac{1}{z}\sqrt{f^2\mathcal{R}^2} \tag{A7}$$

$$V^{im}A^{im} = \frac{1}{z}\sqrt{f^2\mathcal{R}^2 + 2f^2t^2 - 2f^2yt} \tag{A8}$$

$$V^{im}B^{im} = \frac{1}{z}\sqrt{f^2\mathcal{R}^2 + 2f^2t^2 - 2f^2yt} \tag{A9}$$

$$O^{im}V^{im} = \frac{1}{z}\sqrt{f^2t^2} \tag{A10}$$

It is easy to get:

$$O^{im}A^{im} = O^{im}B^{im} \tag{A11}$$

$$V^{im}A^{im} = V^{im}B^{im} \tag{A12}$$

Finally,

$$\begin{aligned} \frac{O^{im}V^{im}}{O^{im}A^{im}} &= \frac{\frac{1}{z}\sqrt{f^2t^2}}{\frac{1}{z}\sqrt{f^2\mathcal{R}^2}} \\ &= \frac{f|t|}{f\mathcal{R}} \\ &= \frac{|t|}{\mathcal{R}} \\ &= |\gamma| \end{aligned} \tag{A13}$$