frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 28 February 2025
DOI 10.3389/fphy.2025.1498335

OPEN ACCESS

EDITED BY
Xin Hu,
East Carolina University, United States

REVIEWED BY
Satyendra Kumar Mishra,
Centre Tecnologic De Telecomunicacions De Catalunya, Spain
Wei Ren,
Xi'an University of Posts and Telecommunications, China

*CORRESPONDENCE
Youbao Chang,
13773583823@163.com

RECEIVED 18 September 2024
ACCEPTED 10 February 2025
PUBLISHED 28 February 2025

CITATION
Qian S, Xue Y and Chang Y (2025) Learning a cross-scale cross-view decoupled denoising network by mining Omni-channel information.
Front. Phys. 13:1498335.
doi: 10.3389/fphy.2025.1498335

Learning a cross-scale cross-view decoupled denoising network by mining Omni-channel information

Song Qian, Yan Xue and Youbao Chang*

Faculty of Information Engineering, Xinjiang Institute of Technology, Aksu, China

Stereo vision systems are increasingly utilized in various applications, however, the presence of noise significantly hampers the quality of the captured images. Traditional denoising methods often fail to address the complex noise patterns in such scenarios, which can adversely affect feature encoding and subsequent processing tasks. This paper introduces a novel stereo denoising approach that leverages cross-view information to enhance the robustness of noise reduction. A Cross-Channel and Spatial Context Information Mining Module is employed to encode long-range spatial dependencies and to bolster inter-channel feature interaction. This module utilizes large convolutional kernels, channel attention mechanisms, and a simple gating structure to enhance feature representation. Our approach relies on an encoder-decoder architecture, which facilitates cross-view and cross-scale feature interactions. The network is trained with a composite loss function that includes both spatial and perceptual domain constraints, ensuring a comprehensive optimization of the denoising process. Extensive experiments conducted on our proposed NoisyST dataset demonstrate the superior performance of our method in terms of noise removal and detail preservation. Notably, the method outperforms existing State-Of-The-Art techniques, as evidenced by its effectiveness in various evaluation metrics.

KEYWORDS

stereo image processing, noise removal, cross-view feature interaction, attentive cross-scale fusion, deeplearning

1 Introduction

In recent years, computer vision tasks have received extensive attention and ushered in rapid development, such as image classification [1, 2], target detection [3], and instance segmentation tasks [4, 5]. These advancements have been significantly propelled by the advent of deep learning, which has revolutionized the way we process and understand visual data. Techniques such as convolutional neural networks (CNNs) have become fundamental in extracting features and making accurate predictions on complex visual tasks. Moreover, the growth of large-scale datasets and the increased computational power have further accelerated the progress in this field. As a result, we are witnessing a transformative era where computer vision systems not only replicate but also, in some cases, surpass human-level performance in various tasks, opening up new possibilities for applications in autonomous driving, medical imaging, and beyond. At the same time, with the wide application of binocular cameras, stereo vision has also ushered in rapid progress. However, binocular

cameras are very sensitive to noise [6], and there is a lack of measures to deal with it. The rapid evolution of deep learning has led to the widespread application of convolutional neural networks (CNNs) in a multitude of single image processing tasks. These include single image deblurring [7], dehazing [8], deraining [9], and enhancement of images under low-light conditions [3]. In parallel, an increasing array of methods leveraging CNNs has been employed in the realm of stereo image processing, further expanding the horizons of what is achievable in the field.

Stereo vision technology has become a cornerstone in various high-precision applications such as robotics, autonomous vehicles, and augmented reality, providing a rich and nuanced understanding of the three-dimensional world. The ability to perceive depth and spatial relationships is paramount for these systems to operate effectively. However, the presence of noise in stereo imagery, often an unavoidable byproduct of real-world imaging conditions, poses a significant challenge to the accuracy and reliability of depth perception. Noise in stereo images not only obscures fine details but also introduces discrepancies between the paired images, which can lead to erroneous depth calculations and subsequent misinterpretations. The traditional noise model is represented as $y = c + n$. Here, y is the noisy image, c is the noise-free image, and n is the noise. This model highlights the need for a denoising strategy. The strategy should effectively remove noise while preserving the image's integrity. While the domain of single image denoising has witnessed remarkable progress with algorithms like DnCNN [10], IRCNN [11], and DRUNet [12], these solutions do not fully translate to the stereo image context. The dual nature of stereo imagery demands a denoising approach that considers the interdependencies and shared information between the two perspectives. Despite advancements in stereo image processing for tasks such as super-resolution and enhancement in low-light conditions, the specific challenge of stereo image denoising has been relatively unaddressed. This paper aims to bridge this gap by introducing an innovative denoising approach that incorporates uncertainty quantification, gradient-based feature enhancement, and a novel frequency interaction mechanism for feature integration.

Our method is designed with the unique characteristics of stereo imagery in mind, focusing on the harmonization of noise reduction and detail preservation across both images of a stereo pair. We demonstrate the efficacy of our approach through rigorous experimentation on a dataset that simulates real-world noise conditions in stereo images. The results indicate that our stereo image denoising technique not only achieves superior noise reduction performance but also maintains the critical details necessary for accurate depth estimation. This work contributes to the literature by offering a robust solution that enhances the resilience of stereo vision systems to noisy conditions, thereby bolstering their applicability in practical scenarios.

In this paper, we propose a robust stereo image enhancement paradigm tailored to address the degradation induced by noise attachment. Furthermore, this paper introduce a strategy aimed at the efficient extraction of interactive interocular information. In summary, the primary contributions of this paper can be categorized into three main aspects:

- NoisyST Dataset for benchmarking deep learning methods in Stereo Image Denoising task. To the best of our knowledge, we are the first to propose a stereo image denoising dataset named NoisyST which contains pairs of clear and noisy stereo images for training and testing neural networks.
- Omni-channel Information Mining Block(OIMB). This paper design a novel module named OIMB for intra-view intra-scale information mining and feature fusion. OIMB can not only realize long-distance modeling, but also effectively capture the information of channel dimension for efficient feature fusion. Firstly, this paper use a Channel-extended Information Mining Module (CIMM) to mine the information flow of the wide-area channel dimension. Secondly, unlike ViT, inspired by NAFNet, this paper propose a mechanism using large convolution kernels to capture long-range dependencies, Large-kernel Long-range Dependency Capture (LLDC).
- Decoupled Infromation Fusion Module(DIFM). This paper design a novel module named DIFM for cross-view cross-scale information mining and feature fusion. DIFM decouples the information fusion into two components, cross-view interactions and cross-scale interactions which focus on view-interaction feature fusion and scale-interaction information minin, respectively.
- CCDDNet: Learning a Cross-Scale Cross-View Decoupled Denoising Network by Mining Omni-channel Information. We propose a network equiped with Omni-channel Information Mining Blocks called CCDDNet. Experiments on NoisyST datasets demonstrate that our proposed framework can recover the details by removing noise in stereo images and obtain SOTA performance.

2 Related work

We briefly review recent progress in single image noise removal and stereo image restoration.

2.1 Single image noise removal

Zhang et al. [10] introduced a deep convolutional neural network (DnCNN) for image denoising that excels in blind Gaussian denoising and extends effectively to tasks like super-resolution and JPEG deblocking, demonstrating high efficiency and performance through residual learning and batch normalization. The work of Zhang et al. [12] presents a significant step forward in the field of image restoration, showcasing the effectiveness of deep learning-based denoising within a flexible Plug-and-Play approach. Zhang et al. [11] have demonstrated the efficacy of integrating fast and powerful CNN denoisers into model-based optimization methods to solve inverse problems in image restoration, showcasing improved performance and flexibility. Thakur and Maji [13] introduced a novel blind denoising approach employing multi-scale pixel attention and feature extraction in a dual-path neural network, demonstrating superior performance with lightweight architecture.

2.2 Stereo image restoration

The iPASSR [14] method utilizes cross-view information and symmetry cues within stereo image pairs, offering a novel and

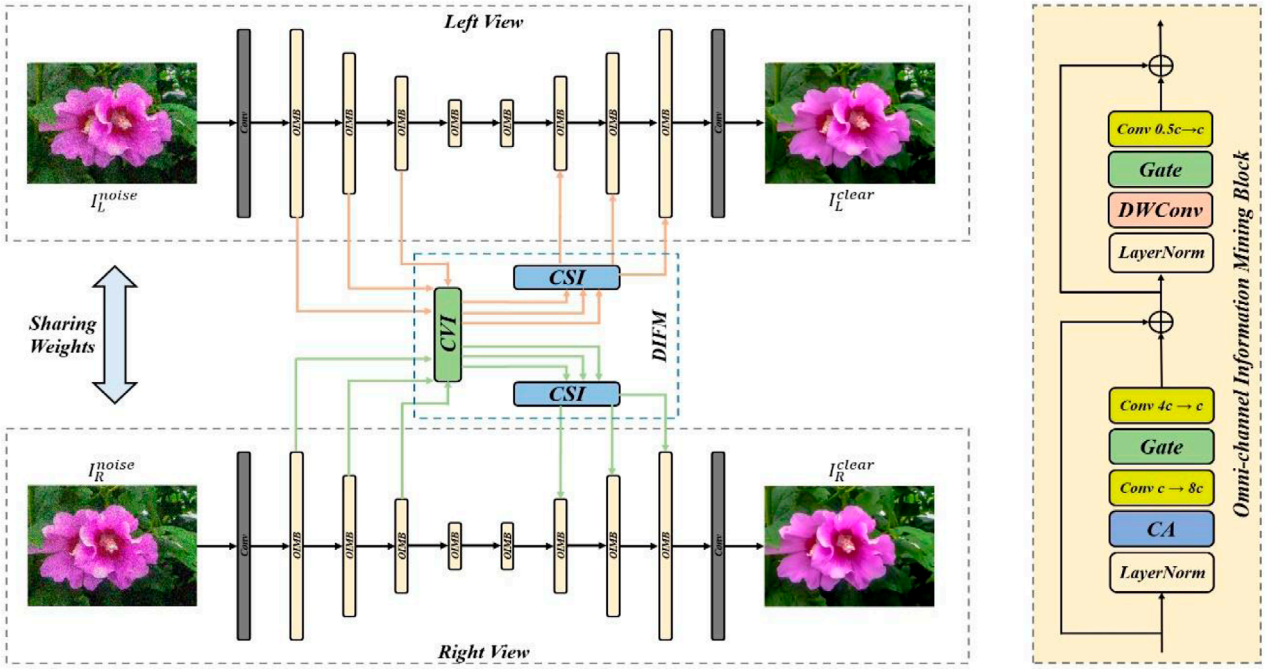

FIGURE 1
The overall framework of our proposed CCDDNet for Stereo Image Denoising. CVI and CSI means cross-view interaction and cross-scale interaction.

effective solution for the challenge of stereo image super-resolution. In the domain of stereo image super-resolution, NAFSSR [15] stands out for its winning performance at the NTIRE 2022 challenge, showcasing the potential of integrating NAFNet's robust feature extraction with cross-view feature fusion. Zheng et al. [6] introduced DCI-Net, a novel approach for low-light stereo image enhancement that leverages decoupled cross-scale cross-view interaction, demonstrating superior performance in illumination adjustment and detail recovery. Zhao et al. [16]introduced a novel approach for low-light stereo image enhancement that simultaneously addresses brightness adjustment and denoising by leveraging a low-frequency information enhanced image space and cross-channel spatial context mining.

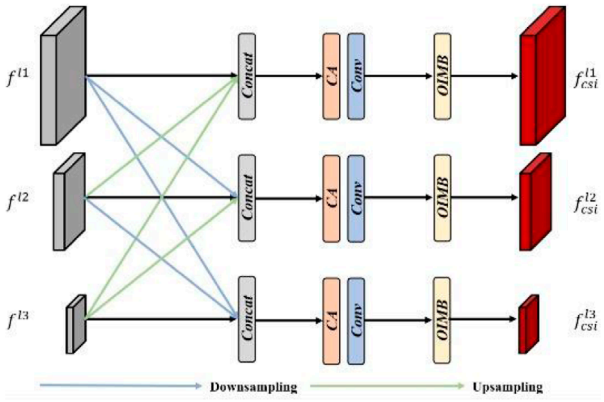

FIGURE 2
The detailed structure of our proposed cross-scale interaction in DIFM. Only the process of left view is shown for simplicity.

3 Proposed framework

In this section, this paper introduce the proposed CCDDNet in detail. We first illustrate the overall architecture of CCDDNet. Then, we describe the individual components of the designed modules, including the Omni-channel Information Mining Block (OIMB) and the Decoupled Information Fusion Module (DIFM). Finally, the used loss functions are discussed.

3.1 Overall structure

The overall structure of our proposed CCDDNet is shown in Figure 1. The proposed model, CCDDNet, is adept at processing a duo of stereo images laden with noise. It adeptly enhances the luminance of each view and subsequently delivers a pair of refined, noise-free stereo images. The methodology unfolds in a three-tiered approach: initial feature extraction from the shallow layers, profound feature extraction from the deeper layers, and ultimately, the reconstruction of the stereo images. Specifically, the model employs a pair of convolutional layers at the onset and terminus of the pipeline. The initial layer is tasked with the extraction of superficial features, whereas the final layer is responsible for the reconstruction of the enhanced, normal-illuminated stereo images. Mathematically, the process for a given set of noisy stereo images can be encapsulated by the following Equation 1:

$$I_L^{clear}, I_R^{clear} = F_{SR}\left(F_{DF}\left(F_{SF}\left(I_L^{noise}, I_R^{clear}\right)\right)\right) \quad (1)$$

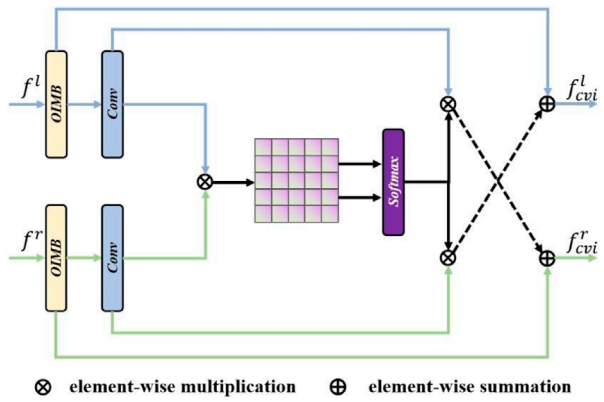

FIGURE 3
The detailed structure of our proposed cross-view interaction in DIFM. Only the process of single scale is shown for simplicity. CVI integrates the information of two perspectives, thus further realizing the complementarity of stereo dimensions and improving the effect of stereo image denoising.

where I_L^{clear}, I_R^{clear}, I_L^{noise} and I_R^{clear} represent the noise-free left-view image, noise-free right-view image, noisy left-view image, noisy right-view image, $F_{SR}(\,)$, $F_{DF}(\,)$ and $F_{SF}(\,)$ denote the transformations of image reconstruction, deep feature extraction and shallow feature extraction, respectively.

3.2 Omni-channel Information Mining Block (OIMB)

In reviewing previous methods for single image denoising and stereo image enhancement, the neural network backbones employed often maintain the channel count within blocks unchanged, followed by the use of residual connections to align inputs with outputs. Such an approach can inadvertently overlook the additional useful auxiliary information encapsulated within the neural network channels. To address this issue, this paper propose a framework called the Channel-wise Information Extraction Module (CIEM), which amplifies the channel count within the network to capture valuable information. Furthermore, Vision Transformers (ViTs) are frequently utilized to capture long-range dependencies and have been demonstrated to effectively enhance accuracy. However, the computational cost associated with ViTs is prohibitively high. Currently, the success of ViTs may be attributed more to the overall architecture rather than self-attention mechanisms [17]. Inspired by this, this paper introduce a Large-kernel Long-range Dependency Capture module (LLDC) designed to seize information on long-range dependencies using large kernel convolutional layer, a technique that has been proven effective [18].

3.2.1 Channel-wise Information Extraction Module

The channel dimension within convolutional neural networks (CNNs) contains a wealth of crucial information that is often underutilized. In the context of stereo image processing, many neural networks maintain a constant channel count in the backbone, which is not conducive to extracting the hidden information in the channel dimension. To address this issue, this paper propose a channel dimension information extraction mechanism. This mechanism employs an exceptionally large channel dimension to expand the original channel eight times and combines attention weights to fully mine the hidden information in the channel dimension of the convolutional neural network. Specifically, the input features first pass through a layer normalization layer to stabilize the distribution, followed by a channel attention layer to empower the features with attention. Then, a 1×1 convolution layer is applied to expand the channel eight times. Next, to enhance the network's ability to extract features and explore the spatial dimension of the feature map, we introduce a gating mechanism that serves as an activation function, reducing the channel count and computational load. Finally, a 1×1 convolution is also used to adjust the channel count back to the original, allowing for the addition of residual connections. This process can be expressed by the following Equation 2:

$$f_{lr} = Conv_{4c\to c}(Gate(Conv_{c\to 8c}(CA(LN(f))))) + f \quad (2)$$

where $\boldsymbol{Conv_{a\to b}}$, *Gate*, $\boldsymbol{CA}$ and $\boldsymbol{LN}$ denote the Conv layer which change the channel from $\boldsymbol{a}$ to $\boldsymbol{b}$, a simple gate module, channel attention and layer normalizaton, respectively. $\boldsymbol{f}$ and $\boldsymbol{f_{lr}}$ denote the input and output feature maps, respectively.

3.2.2 Large-kernel long-range Dependency Capture module

Divergent from the Channel Information Extraction Module (CIEM) introduced in the initial phase of the OIMB, which focuses on channel dimension information, the Large-kernel Long-range Dependency Capture (LLDC) stage primarily delves into the extraction of spatial dimension information. Echoing the foundational premise of the visual Transformer, our objective in the LLDC stage is to seize long-distance dependencies within the data. Concurrently, to mitigate computational overhead, this paper opt for a large convolution kernel to fulfill this objective efficiently. In the LLDC stage, we employ a large kernel size to cover a broader spatial context and capture the intricate patterns that span across wider regions of the image. This approach allows us to tap into the rich, long-range spatial correlations that are often crucial for understanding the global structure of scenes within visual tasks. By doing so, we enhance the model's ability to recognize coherent objects and shapes, which is particularly beneficial for tasks like image segmentation and object detection where holistic scene understanding is required. Moreover, the use of a single, large kernel also helps to reduce the number of parameters and computations compared to a stack of smaller convolutions, thus striking a balance between the model's complexity and its performance. This strategic design choice ensures that our model remains lightweight and efficient, making it suitable for real-time applications and devices with limited computational resources. As depicted in Figure 1, the LLDC process is succinctly encapsulated in the following Equation 3:

$$f_{lr} = Conv_{0.5c\to c}(Gate(DWConv_k(LN(f)))) + f \quad (3)$$

where $\boldsymbol{DWConv_k}$ means the employed large-kernel convolution layer with kernel size k.

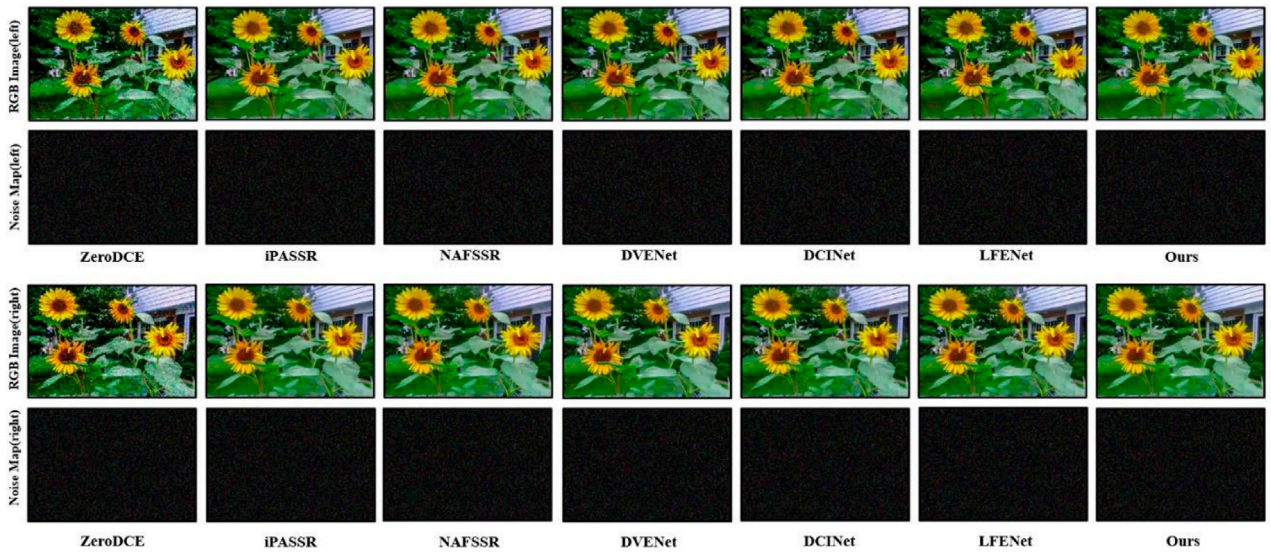

FIGURE 4
Visual results of SOTA methods on Flickr1024 dataset.

FIGURE 5
Visual results of SOTA methods on Kitti2012 dataset.

3.3 Decoupled Information Fusion Module (DIFM)

Stereo image processing is different from monocular image processing. One of its key points is to explore the correlation between the two views in order to better extract features, which is particularly evident in the task of stereo image denoising. Therefore, since most monocular image denoising methods only consider one view [10, 11], they are not effective in enhancing stereoscopic images. Although some existing algorithms for stereo image processing have mastered the information interaction between cross-views [14, 15], they lack the understanding of the importance of cross-scale information interaction. However, the importance of cross-scale information interaction is important in stereo image processing [6]. In order to solve the above problems, this paper propose a Decoupled Information Fusion Module (DIFM), which decouples cross-scale information and cross-view information, and studies the importance of the two to promote further feature fusion and interaction.

3.3.1 Cross-scale interaction

The integration of cross-scale information is crucial for enhancing the performance of Stereo image denoising tasks. Despite its significance, many existing Stereo image processing

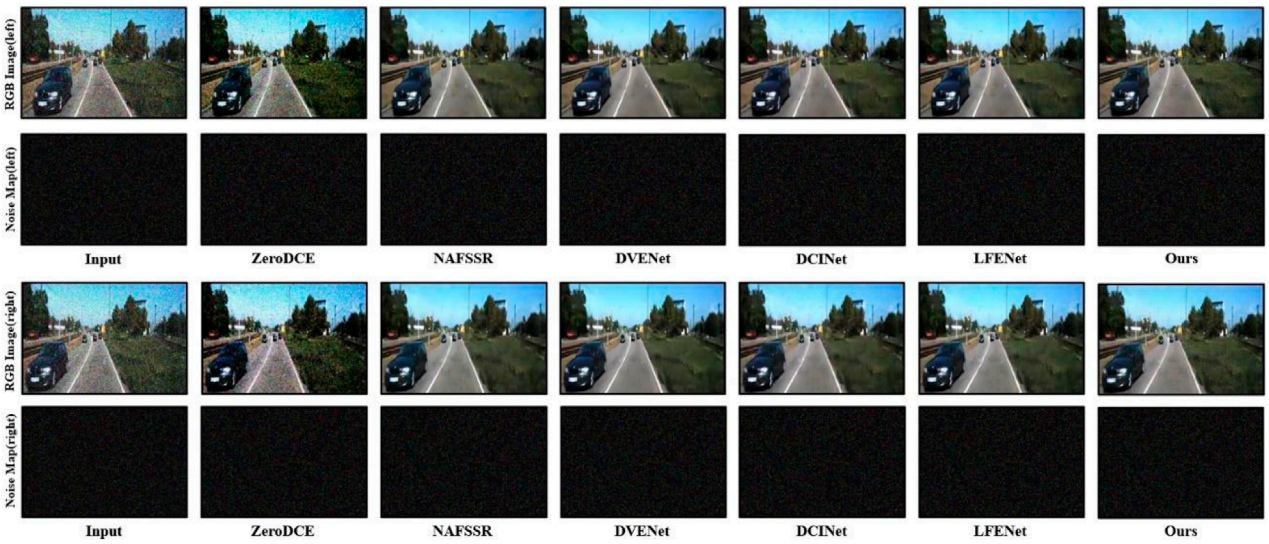

FIGURE 6
Visual results of SOTA methods on Kitti2015 dataset.

TABLE 1 Comparative results on synthetic stereo noisy images and the noise level is 15.

Method	Venue	Flickr1024		Kitti2012		Kitti2015		Middlebury	
		Left	Right	Left	Right	Left	Right	Left	Right
DnCNN	TIP'17	25.47/0.916	21.70/0.810	22.29/0.906	19.15/0.803	19.03/0.890	16.63/0.790	19.22/0.892	16.78/0.793
DRUNet	TPAMI'21	30.51/0.940	30.55/0.940	30.61/0.942	30.67/0.942	30.63/0.943	30.74/0.943	30.65/0.943	30.77/0.944
IRCNN	CVPR'17	30.52/0.939	30.54/0.939	30.61/0.940	30.65/0.941	30.64/0.941	30.72/0.942	30.66/0.941	30.74/0.943
NGDCNet	Electronics'23	27.46/0.912	26.90/0.907	27.91/0.917	27.40/0.914	28.40/0.922	27.99/0.919	28.55/0.924	28.16/0.921
NIFBGDNet	PR'23	31.41/0.950	31.44/0.951	31.46/0.951	31.52/0.952	31.46/0.951	31.57/0.952	31.48/0.952	31.60/0.953
DVENet	TMM'22	31.46/0.952	31.47/0.952	31.49/0.952	31.53/0.953	31.46/0.952	31.56/0.953	31.50/0.953	31.60/0.953
NAFSSR	CVPRW'22	31.59/0.951	31.58/0.951	31.63/0.952	31.63/0.952	31.61/0.952	31.68/0.952	31.64/0.952	31.71/0.953
iPASSR	CVPWW'21	31.40/0.949	31.40/0.949	31.45/0.950	31.47/0.950	31.45/0.950	31.53/0.951	31.49/0.951	31.57/0.951
LFENet	Arxiv'24	29.99/0.922	29.99/0.922	30.22/0.927	30.25/0.927	30.30/0.930	30.40/0.931	30.36/0.931	30.47/0.932
DCINet	MM'23	30.70/0.948	30.44/0.946	30.78/0.949	30.57/0.947	30.77/0.949	30.64/0.948	30.80/0.950	30.67/0.949
Ours	—	31.72/0.955	31.71/0.954	31.75/0.955	31.76/0.955	31.73/0.955	31.80/0.955	31.76/0.956	31.84/0.956

techniques overlook this aspect. To address this oversight, this paper have developed a solution that adeptly and efficiently merges cross-scale data, thereby enhancing the quality of feature fusion. The architecture of this cross-scale interaction is depicted in Figure 2. As the left view features $f_{csi}^{l_1}, f_{csi}^{l_2}, f_{csi}^{l_3}$ of the input stereo image are presented initially, the process begins with the alignment of features across different scales, followed by their seamless integration. Subsequently, a 1×1 convolution is applied to condense the channel dimensions and facilitate inter-channel communication. To optimize the utilization of each channel's distinctive attributes, this paper incorporate a channel attention mechanism prior to the convolutional operation. Concluding the sequence, an OIMB is introduced to foster deeper feature interaction and extraction. The methodology aforementioned is encapsulated in the following Equation 4:

$$\mathrm{f}_{\mathrm{csi}}^{\mathrm{l}_1}, \mathrm{f}_{\mathrm{csi}}^{\mathrm{l}_2}, \mathrm{f}_{\mathrm{csi}}^{\mathrm{l}_3} = \mathrm{OIMB}\left(\mathrm{Gate}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{FC}\left(\mathrm{f}^1, \mathrm{f}^2, \mathrm{f}^3\right)\right)\right)\right) \tag{4}$$

where $FC()$ denotes the feature concatenation operation in the Cross-view interaction, $f_{csi}^{l_1}, f_{csi}^{l_2}, f_{csi}^{l_3}$ denote the output feature maps.

TABLE 2 Comparative results on synthetic stereo noisy images and the noise level is 25.

Method	Venue	Flickr1024		Kitti2012		Kitti2015		Middlebury	
		Left	Right	Left	Right	Left	Right	Left	Right
DnCNN	TIP'17	27.06/0.896	20.95/0.711	26.70/0.897	20.76/0.713	25.57/0.894	20.08/0.711	25.69/0.895	20.13/0.713
DRUNet	TPAMI'21	28.78/0.911	28.52/0.906	28.88/0.913	28.61/0.909	28.91/0.914	28.69/0.911	28.95/0.915	28.72/0.913
IRCNN	CVPR'17	28.01/0.898	28.05/0.899	28.15/0.901	28.17/0.902	28.20/0.903	28.27/0.904	28.30/0.906	28.26/0.905
NGDCNet	Electronics'23	27.18/0.900	26.98/0.896	27.57/0.905	27.40/0.902	27.81/0.908	27.72/0.906	27.90/0.909	27.81/0.907
NIFBGDNet	PR'23	28.82/0.914	28.83/0.931	28.83/0.914	28.92/0.916	28.96/0.917	29.02/0.917	28.98/0.918	29.06/0.918
DVENet	TMM'22	29.07/0.921	29.08/0.921	29.16/0.922	29.18/0.923	29.18/0.923	29.25/0.924	29.21/0.924	29.29/0.925
NAFSSR	CVPRW'22	28.89/0.915	28.88/0.915	28.98/0.917	28.98/0.917	29.01/0.918	29.06/0.918	29.04/0.919	29.09/0.919
iPASSR	CVPWW'21	28.83/0.913	28.83/0.913	28.83/0.915	28.94/0.916	28.96/0.917	29.03/0.917	29.00/0.918	29.06/0.918
LFENet	Arxiv'24	27.31/0.902	27.35/0.902	27.54/0.906	27.58/0.906	27.65/0.908	27.75/0.909	27.65/0.909	27.77/0.910
DCINet	MM'23	28.62/0.918	28.55/0.917	28.74/0.919	28.67/0.919	28.77/0.920	28.77/0.902	28.80/0.921	28.80/0.921
Ours	—	29.28/0.925	29.27/0.924	29.34/0.926	29.34/0.926	29.35/0.927	29.41/0.927	29.39/0.928	29.45/0.928

TABLE 3 Comparative results on synthetic stereo noisy images and the noise level is 50.

Method	Venue	Flickr1024		Kitti2012		Kitti2015		Middlebury	
		Left	Right	Left	Right	Left	Right	Left	Right
DnCNN	TIP'17	23.39/0.743	15.67/0.496	23.47/0.749	15.68/0.501	23.53/0.754	15.68/0.504	23.54/0.756	15.69/0.506
DRUNet	TPAMI'21	25.01/0.812	25.03/0.812	25.16/0.818	25.18/0.818	25.24/0.822	25.31/0.823	25.26/0.825	25.33/0.826
IRCNN	CVPR'17	24.84/0.801	24.83/0.798	24.99/0.808	24.97/0.805	25.06/0.814	25.07/0.811	25.08/0.816	25.18/0.813
NGDCNet	Electronics'23	24.91/0.826	24.66/0.823	25.17/0.832	24.93/0.830	25.33/0.838	25.15/0.836	25.38/0.840	25.21/0.838
NIFBGDNet	PR'23	25.54/0.828	25.64/0.832	25.67/0.833	25.77/0.837	25.74/0.837	25.88/0.842	25.76/0.840	25.91/0.844
DVENet	TMM'22	25.86/0.842	25.88/0.843	26.00/0.846	26.00/0.848	26.06/0.850	26.11/0.852	26.09/0.852	26.15/0.854
NAFSSR	CVPRW'22	25.38/0.825	25.39/0.825	25.51/0.830	25.51/0.830	25.58/0.834	25.62/0.834	25.61/0.836	25.65/0.836
iPASSR	CVPWW'21	25.62/0.827	25.62/0.828	25.75/0.832	25.75/0.833	25.81/0.836	25.85/0.838	25.84/0.838	25.89/0.840
LFENet	Arxiv'24	25.37/0.833	25.38/0.832	25.56/0.838	25.56/0.838	25.53/0.840	25.57/0.841	25.57/0.843	25.63/0.843
DCINet	MM'23	25.56/0.841	25.55/0.840	25.78/0.845	25.68/0.844	25.76/0.848	25.79/0.848	25.78/0.850	25.82/0.850
Ours	—	26.05/0.853	26.07/0.853	26.16/0.856	26.16/0.856	26.20/0.859	26.25/0.859	**26.22/0.861**	26.28/0.861

3.3.2 Cross-scale interaction

Incorporating the aforementioned cross-scale information fusion, this paper address the prevalent issue of insufficient cross-scale integration in most stereo-view interaction methodologies by proposing a multi-scale stereo-view fusion approach. This innovative multi-scale stereo-view fusion approach systematically integrates information across different scales to enhance the interaction between stereo views. By leveraging a pyramid structure, we enable the model to capture both local details and global context effectively. At each scale, the stereo views are first processed independently to extract features, and then fused using our cross-scale information fusion strategy. This fusion operation allows the model to leverage the complementary strengths of both views, such as depth cues and texture information, which are critical for tasks like depth estimation and scene understanding in stereo images. The multi-scale nature of our approach also ensures robustness against

TABLE 4 Ablation studies on the effects of designed backbones, losses, and modules of our proposed method. The noise level is 15.

Method	Left		Right		Mean	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
W/o CIEM	28.59	0.916	28.60	0.915	28.60	0.915
W/o LLDC	28.80	0.918	28.81	0.917	28.81	0.918
W/o CVI	28.45	0.915	28.47	0.915	28.46	0.915
W/o CSI	28.50	0.912	28.51	0.912	28.50	0.912
W/o DIFM	28.32	0.912	28.32	0.911	28.32	0.912
W/o L_{char}	27.43	0.888	27.39	0.889	27.41	0.889
W/o L_{perp}	28.69	0.917	28.71	0.917	28.70	0.917
Ours	29.28	0.925	29.27	0.924	29.27	0.925

various levels of noise and occlusions, which are common challenges in real-world stereo vision applications. The intricate structure of this method is illustrated in Figure 3. The input is a stereo feature map. Our Cross-View Interaction (CVI) mechanism computes an interaction weight matrix, which is then utilized to refine the synthesis of features across different views. This sophisticated process is succinctly articulated by the subsequent Equation 5:

$$W_m = f^l(f^r)^T \tag{5}$$

where $\boldsymbol{f^l}$ and $(\boldsymbol{f^r})^T$ denote the input stereo feature maps, $\boldsymbol{W_m}$ denotes the weight of the fusion matrix.

3.4 Loss function

The total loss function $\boldsymbol{L}$ used in this paper contains two losses, i.e., perceptual loss [19] and Charbonnier loss [20]. The Charbonnier loss can be formulated as:

$$L_{char} = \sqrt{\left\| I^{out} - \bar{I} \right\|_2 + \epsilon^2}$$

where $\bar{\boldsymbol{I}}$ and $\boldsymbol{I^{out}}$ are the ground truth and output of the whole network. ϵ is set to 0.001. The perceptual loss compares the VGG-19 [21] feature distances between $\bar{\boldsymbol{I}}$ and $\boldsymbol{I^{out}}$ using an L1 loss as:

$$L_{perp} = \left\| \emptyset(I^{out}) - \emptyset(\bar{I}) \right\|_1$$

where $\emptyset()$ denotes the feature extraction operation from VGG-19 network [21]. Therefore, the total loss in this paper is:

$$L = L_{char} + \alpha L_{perp}$$

where α is a hyper parameter.

3.5 Experiment settings

This paper propose the first binocular image denoising dataset NoisyST for training neural networks. Following the previous works [6, 14], this paper also select data from the existing stereo image dataset to synthesize. Specifically, for the training dataset, we utilize 800 stereo image pairs from Flickr1024 [22] dataset and 60 stereo image pairs from Middlebury [23] dataset; for the testing set, this paper select 112 stereo image pairs from Flickr1024 [22], 20 stereo image pairs from KITTI2012 [24], 5 stereo image pairs from Middlebury [23] and 20 stereo image pairs from KITTI2015 dataset [25]. For the synthesis of noisy images, this paper use the method of randomly adding noise during the training process. For the evaluation method, this paper use the widely used PSNR and SSIM. The higher the two indicators are, the better the image quality is.

All experiments are conducted by using Pytorh with NVIDIA RTX3080GPUs. Following [6], we crop images into 128×128 pixels. This paper employ AdamW optimizer with a mini-batch 4, the initial learning rate is 2×10^{-4} and this paper use CosineAnnealingLR for weight decay. We train our CCDDNet for 400 epochs. For comparing SOTA methods, this paper utilize DnCNN [10], IRCNN [11], DRUNet [12], NGDCNet [26], NIFBGDNet [13], iPASSR [14], DVENet [27], NAFSSR [15], SNR [28], DCINet [6] and LFENet [16].

4 Experimental results and analysis

Firstly, we introduce the experimental setting in this paper. Then, we show the experimental results and analysis of our proposed method.

4.1 Visual results

To enhance clarity and visual appeal, this paper present the denoised stereo images alongside their noisy counterparts in Figures 4, 6, showcasing various components of our NoisyST dataset, including ZeroDCE [29], iPASSR [14], NAFSSR [15], DVENet [27], DCINet [6], LFENet [16].

Figure 4 illustrates the superior noise reduction capabilities of our method on the Flickr2014 stereo image dataset, resulting in the highest quality image restoration. Figures 5, 6 further demonstrate the effectiveness of our method in the context of autonomous driving, highlighting the impact of denoising on the KITTI2012 and KITTI2015 datasets. It is evident that our approach not only recovers the clearest stereo images but also detects the most noise, significantly enhancing autonomous driving perception and subsequent processes.

4.2 Quantitative results

For our proposed CCDDNet, this paper evaluate its performance on our proposed NoisyST dataset at three different noise levels. Tables 1–3 shows the numerical results of left view and right view of denoising images at different noise levels, e.g., 15, 25 and 50. Tables 1–3 clearly demonstrates that our proposed method excels in recovering images with superior visual quality and detecting noise more effectively than other methods. It is evident that methods dedicated to single image denoising, such as DnCNN [10], DRUNet [12], IRCNN [11], NGDCNet [26], NIFBGDNet [13], are limited by their lack of robust mechanisms for information interaction, which hinders their ability to effectively integrate the left and right perspectives in stereo images. Moreover, stereo image processing methods, including DVENet [27], NAFSSR [15], and iPASSR [14], are found to be deficient in cross-view multi-scale information interaction. The cross-view interaction techniques they employ are also suboptimal. Lastly, methods that utilize cross-view and cross-scale information interaction, such as LFENet [16] and DCINet [6], require refinement of their information interaction techniques. Concurrently, the foundational backbone networks they employ are lacking in their capacity to model channel dimension information and capture long-range dependencies.

4.3 Ablation studies

This paper show the ablation studies to demonstrate the rationality of independent components of our proposed CCDDNet, including designed modules, losses and backbones. The ablation experiments are performed on the Flickr1024 part of our proposed NoisyST dataset. The numerical results are shown in Table 4.

4.3.1 Effectiveness of OIMB

To demonstrate the effectiveness of our proposed OIMB, this paper use two modules as shown in Table 4. Specifically, W/o CIEM and W/o LLDC denote removing CIEM and LLDC from OIMB respectively. From Table 4, it clearly illustrates that the capabilities of CIEM and LLDC to independently extract channel dimension information and capture long-distance dependencies, respectively, are crucial for optimal performance. It is evident that the removal of either component leads to a significant decline in performance, underscoring the indispensable nature of both mechanisms in maintaining the module's effectiveness.

4.3.2 Effectiveness of DIFM

This paper subsequently confirm the contribution of DIFM by dropping CVI, CSI and DIFM in our proposed CCDDNet, which are denoted as W/o CVI, W/o CSI and W/o DIFM in Table 4. It can be seen from the table that the removal of CVI and CSI alone will cause a serious reduction in performance. Further, when the entire DIFM is removed, it will cause the most serious reduction in performance, which indicates the rationality of using DIFM and the excellent performance of DIFM.

4.3.3 Effectiveness of losses

We proceed to confirm the role of the loss functions in this paper. As shown in Table 4, W/o L_{char} and W/o L_{perp} means removing L_{char} and L_{perp} while training the model. The table reveals that the omission of a loss function invariably leads to diminished performance. This finding affirms the logic of utilizing a composite loss function strategy, which integrates multiple components to optimize the model's performance.

5 Conclusion

This paper explore a new vision task, stereo image denoising, and we propose a new benchmark called NoisyST dataset which can be used for training and testing neural networks. In general, this paper propose a novel model for stereo image denoising, called CCDDNet. Specifically, addressing the deficiency in effective cross-view information interaction and cross-scale information fusion within stereo vision image processing tasks, this paper delve into the development of robust solutions tailored for stereo image denoising, thereby enhancing the overall performance of stereo image denoising techniques. Further, aiming at the lack of feature extraction ability of channel dimension and the ability of long-distance dependence capture in stereo image denoising, this paper propose a backbone network module called OIMB, including channel dimension information mining module CIEM and long-distance dependence capture module LLDC. These two modules are responsible for mining channel dimension information and capturing long-distance dependencies in the network learning process. The comparative experiments conducted on our NoisyST dataset demonstrate not only its suitability as a benchmark for training neural networks but also the exceptional performance of our proposed CCDDNet. Our method stands out in restoring images with the highest visual quality and achieving the most outstanding results. Additionally, the ablation study of CCDDNet's components further validates the soundness of our approach. In the future, we are committed to exploring even more efficient methods for stereo image denoising.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SQ: Conceptualization, Funding acquisition, Software, Validation, Visualization, Writing–original draft, Writing–review

and editing. YX: Data curation, Supervision, Validation, Writing–review and editing. YC: Conceptualization, Methodology, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was Sponsored by Xinjiang Uygur Autonomous Region Tianshan Talent Programme Project (No.2023TCLJ02).

Acknowledgments

We thank all the editors and reviewers in advance for their valuable comments that will improve the presentation of this paper.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Iqbal I, Odesanmi GA, Wang J, Liu L. Comparative investigation of learning algorithms for image classification with small dataset. *Appl Artif Intelligence* (2021) 35(10):697–716. doi:10.1080/08839514.2021.1922841

2. He K, Zhang X, Ren S. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer 345 vision and pattern recognition* (2016). p. 770–8.

3. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: *IEEE transactions on pattern analysis and machine intelligence* (2016) 39(6), 1137–1149. doi:10.1109/TPAMI.2016.2577031

4. Iqbal I, Shahzad G, Rafiq N, Mustafa G, Ma J. Deep learning-based automated detection of human knee joint's synovial fluid from magnetic resonance images with transfer learning. *IET Image Process* (2020) 14(10):1990–8. doi:10.1049/iet-ipr.2019.1646

5. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans pattern Anal machine intelligence* (2017) 39(12):2481–95. doi:10.1109/tpami.2016.2644615

6. Zheng H, Zhang Z, Fan J. Decoupled cross-scale cross-view interaction for stereo image enhancement in the dark. In: *Proceedings of the 31st ACM international conference on multimedia* (2023). p. 1475–84.

7. Nah S, Hyun Kim T, Mu Lee K. Deep multi-scale convolutional neural network for dynamic scene deblurring. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017). p. 3883–91.

8. Qin X, Wang Z, Bai Y, Xie X, Jia H. FFA-Net: feature fusion attention network for single image dehazing. In: *Proceedings of the AAAI conference on artificial intelligence* (2020) 34(07):11908–15. doi:10.1609/aaai.v34i07.6865

9. Jiang K, Wang Z, Yi P. Multi-scale progressive fusion network for single image deraining. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020). p. 8346–55.

10. Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a Gaussian denoiser: residual learning of deep cnn for image denoising. *IEEE Trans image Process* (2017) 26(7):3142–55. doi:10.1109/tip.2017.2662206

11. Zhang K, Zuo W, Gu S, Zhang L. Learning deep CNN denoiser prior for image restoration. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017). p. 3929–38.

12. Zhang K, Li Y, Zuo W, Zhang L, Van Gool L, Timofte R. Plug-and-play image restoration with deep denoiser prior. *IEEE Trans Pattern Anal Machine Intelligence* (2021) 44(10):6360–76. doi:10.1109/tpami.2021.3088914

13. Thakur RK, Maji SK. Multi scale pixel attention and feature extraction based neural network for image denoising. *Pattern Recognition* (2023) 141:109603. doi:10.1016/j.patcog.2023.109603

14. Wang Y, Ying X, Wang L. Symmetric parallax attention for stereo image super-resolution. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021). p. 766–75.

15. Chu X, Chen L, Yu W. Nafssr: stereo image super-resolution using nafnet. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). p. 1239–48.

16. Zhao M, Qin X, Du S, Bai X, Lyu J, Liu Y. Low-light stereo image enhancement and de-noising in the low-frequency information enhanced image space. *Expert Systems with Applications* (2025). 265, 125803. doi:10.1016/j.eswa.2024.125803

17. Yu W, Luo M, Zhou P. Metaformer is actually what you need for vision. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). p. 10819–29.

18. Ding X, Zhang X, Han J. Scaling up your kernels to 31x31: revisiting large kernel design in cnns. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). p. 11963–75.

19. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. In: Proceedings, Part II 14 Computer Vision–ECCV 2016: 14th European Conference; October 11-14, 2016; Amsterdam, The Netherlands. Springer International Publishing (2016). p. 694–711.

20. Lai WS, Huang JB, Ahuja N, Yang MH. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Trans pattern Anal machine intelligence* (2018) 41(11):2599–613. doi:10.1109/tpami.2018.2865304

21. Simonyan K. *Very deep convolutional networks for large-scale image recognition* (2014).

22. Wang L, Wang Y, Liang Z. Learning parallax attention for stereo image super-resolution. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019). p. 12250–9.

23. Scharstein D, Hirschmüller H, Kitajima Y. High-resolution stereo datasets with subpixel-accurate ground truth. In: Proceedings 36 Pattern Recognition: 36th German conference, GCPR 2014, Münster, Germany. September 2-5, 2014: Springer Interna 383 tional Publishing (2014). p. 31–42.

24. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving the kitti vision benchmark suite. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE (2012). p. 3354–61.

25. Menze M, Geiger A. Object scene flow for autonomous vehicles. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015). p. 3061–70.

26. Zhu M, Li Z. NGDCNet: noise gating dynamic convolutional network for image denoising. *Electronics* (2023) 12(24):5019. doi:10.3390/electronics12245019

27. Huang J, Fu X, Xiao Z, Zhao F, Xiong Z. Low-light stereo image enhancement. *IEEE Trans Multimedia* (2022) 25:2978–92. doi:10.1109/tmm.2022.3154152

28. Xu X, Wang R, Fu CW. SNR-aware low-light image enhancement. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). p. 17714–24.

29. Guo C, Li C, Guo J. Zero-reference deep curve estimation for low-light image enhancement. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020). p. 1780–9.