Dynamic Grouped Interaction Network for Low-Light Stereo Image Enhancement

Baiang Li
Hefei University of Technology
Hefei, China

Huan Zheng
Hefei University of Technology
Hefei, China

Zhao Zhang
Hefei University of Technology
Hefei, China

Yang Zhao
Hefei University of Technology
Hefei, China

Zhongqiu Zhao
Hefei University of Technology
Hefei, China

Haijun Zhang
Harbin Institute of Technology (Shenzhen)
Shenzhen, China

ABSTRACT

Low-Light Stereo Image Enhancement (LLSIE) tackles the challenge of improving the illumination and restoring the details in stereo images. However, existing deep learning-based LLSIE methods trained on high-resolution low-light images often exhibit sub-optimal performance when interacting with information from the left and right views. We find that this is because of: (1) the high computational cost arising from quadratic complexity, which hinders the enhancement model's ability to process high-resolution images; and (2) the limitations of conventional fusion strategies in previous work, which inadequately capture cross-view cues, resulting in weak feature representation and compromised detail recovery. To address these limitations, we propose a novel Dynamic Grouped Interaction Network (DGI-Net) to enhance illumination and recover more details while reducing the computational cost. Specifically, DGI-Net employs the U-Net structure, which effectively mitigates noise during the low-light enhancement. Furthermore, we design a Grouped Stereo Interaction Module (GSIM) with a grouping strategy to efficiently discover cross-view cues while minimizing computations. To dynamically fuse stereo information and fully exploit cross-view correlations, we also introduce a Dynamic Embedding Module (DEM) to establish dynamic connections between inter-view cues and intra-view features, which performs dynamic weight processing on cross-view cues to eliminate noise during fusion. For intra-view processing, we present a Diversity Enhanced Block (DEB) to extract multi-scale features, thereby improving diversity and feature representation. This multi-scale feature extraction also addresses low image contrast in dark lighting conditions. Experimental results demonstrate that DGI-Net outperforms current state-of-the-art methods in low-light stereo image enhancement.

CCS CONCEPTS

• **Computing methodologies** → **Image representations; Stereo image enhancement**.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611895

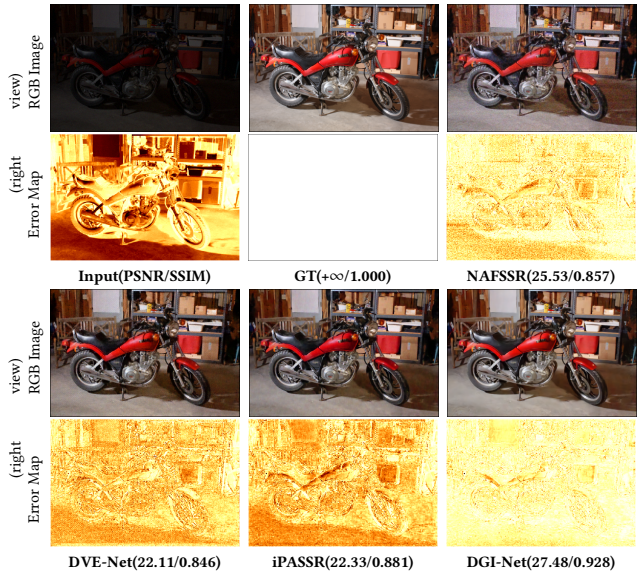

Figure 1: Visual comparison results of DGI-Net with NAFSSR [2] , DVE-Net [8] and iPASSR [31] on Middlebury dataset. In each group of results, the top image represent the reconstructed images, while the bottom one denotes the error map [10] in comparison to the ground truth. Specifically, these error maps are designed to represent data with a clear progression from low to high values. The error maps utilize a black-to-white gradient, which passes through dark red, orange, and yellow colors. The gradient is then reversed, resulting in the error maps where less visible content indicates that the result is closer to the ground truth. As can be observed, the enhanced result of our DGI-Net recovers more illumination details and introduce less redundant noise, owing to the dynamic information interaction process.

KEYWORDS

Low-level vision, low-light image enhancement, grouping strategy, dynamic convolution, diversity enhanced block

ACM Reference Format:
Baiang Li, Huan Zheng, Zhao Zhang, Yang Zhao, Zhongqiu Zhao, and Haijun Zhang. 2023. Dynamic Grouped Interaction Network for Low-Light Stereo Image Enhancement. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3611895

1 INTRODUCTION

Image processing and understanding [32, 42, 33, 27, 26, 44, 40, 43] have gained significant attention in the recent decades. Recently, the popularization of stereo camera equipment has led to increased interests in stereo image processing. However, due to unavoidable lighting conditions, numerous low-light images are captured [18],

which negatively affects the aesthetic quality. These images often exhibit unclear content and texture, low contrast and unknown noise, which in turn hinders subsequent computer vision tasks. To address these issues, researchers have focused on enhancing the low-quality images captured in dark conditions. For single images, Low-Light Image Enhancement (LLIE) methods [21, 1, 17, 22, 11] have been developed, which build models to refine lighting and produce enhanced images. These methods primarily include histogram equalization-based [21] [1] and retinex-based [11] techniques. Although traditional LLIE methods are relatively simple and easy to implement [7, 3, 9, 16], the enhanced results often exhibit unpleasant details.

In the context of stereo images, researchers have turned their attention to stereo image restoration (SIR) [2, 31, 12, 34, 36] which enhances the quality of degraded stereo images by utilizing complementary information provided by binocular systems. Low-Light Stereo Image Enhancement (LLSIE) has also been proposed to refine the illumination and improve the visual quality of stereo images with poor lighting conditions to obtain natural, normal-light images. A key aspect of SIR methods, and LLSIE in particular, is discovering and exploiting cross-view relationships to recover lost information. Existing methods have used shift operations and parallax attention to explore cues between stereo pairs [31] [12]. However, the restored results still suffer from inaccurate illumination and details, as shown in Figure 1. This is because the LLSIE approach still faces the following three challenges:

- **High computational cost.** Stereo information interaction typically involves matrix multiplication on the entire feature maps of both views. This approach is suitable for low-resolution image restoration but is too computationally expensive for high-resolution LLSIE task.
- **Inefficient information fusion.** After obtaining cues between left and right views, existing works [31] [8] [2] perform information fusion using element-wise addition or other simple convolution modules. These methods only perform basic fusion of cross-view cues and intra-view features without considering whether cross-view cues consistently benefit image detail recovery. As a result, cross-view cues may introduce redundant noise during fusion, negatively impacting image detail recovery.
- **Inaccurate feature representation.** Current methods primarily focus on exploring cross-view relationship, with less attention paid to intra-view processing. In addition, solutions that utilize feature extraction blocks for intra-view feature extraction often neglect improving feature diversity, resulting in inaccurate feature representation, high noise and low contrast during enhancement, further compromising enhanced image quality.

To address these limitations, we propose a grouping strategy to reduce the computational cost and recover more image details by filtering effective features and extracting features at multiple scales. Besides, the proposed method effectively enhances illumination and recovers hidden details for LLSIE. The main contributions of this paper are summarized as follows:

- **Dynamic Grouped Interaction Network (DGI-Net)**. We propose DGI-Net for LLSIE, which offers novel strategies for improving both inter-view and intra-view processing. The architecture of our model can better remove the noise during the process of LLSIE. Additionally, it fully interacts with stereoscopic view information, filters out effective features for image detail recovery, and extracts multi-scale features of inter-view and intra-view to address low contrast in low-light images. Extensive experiments on Flickr1024, KITTI 2012, KITTI 2015 and Middlebury datasets demonstrate that our model delivers better illumination correction and texture restoration compared to other related methods.
- **Grouped Stereo Interaction Module (GSIM)**. To reduce the computational cost, we present GSIM with a novel grouping strategy. Instead of using matrix multiplication on the entire feature maps to capture cross-view cues as in previous works, GSIM groups feature maps and employs subsets to discover relationships between stereo pairs. As a result, GSIM can complete stereo information interaction with linear computational complexity.
- **Dynamic Embedding Module (DEM)**. We design DEM to address the issue of inefficient information fusion and the under-utilization of cross-view cues. Specifically, We introduce a dynamic fusion strategy to build dynamic connections between cross-view cues and intra-view features. This dynamic mechanism effectively filters out useful cross-view features and integrate them into the original feature maps reasonably, facilitating the full use of stereo cues and inter-view information flow.
- **Diversity Enhanced Block (DEB)**. To further process intra-view features and address the problem of low contrast and high noise in low-light images, DEB incorporates multiple convolution layers with different kernel sizes to extract multi-scale features. These multi-scale features are then fused to improve the diversity of information, obtain better feature representation, and further refine the enhanced results. Also, this multi-scale detail recovery process can effectively suppress noise generated during the low-light image enhancement process.

2 RELATED WORK

Traditional image enhancement methods [7, 3, 30, 5, 20] can improve the brightness to a certain extent, but the restored results often suffer from over- or under-enhancement. As such, in this section, we focus on the deep-learning-based single LLIE methods and stereo image restoration methods.

2.1 Single Low-Light Image Enhancement

Single low-light image enhancement aims to enhance illumination and recover image details captured in low-light conditions. Similar to other single image restoration and enhancement tasks, the U-Net [24] framework is widely used as the backbone. Jiang et al. [13] designed a U-shape 3D network to learn the mapping from raw low-light videos to normal-light videos. EnlightenGAN [14] proposed an attention-guided U-Net as the generator with a global-local discriminator to enhance low-light images without paired data. To further reduce reliance on the used data, Li et al. [19] proposed a deep curve estimation method in a zero-shot manner. Recently, researchers have explored other methods. Xu et al. [35] presented

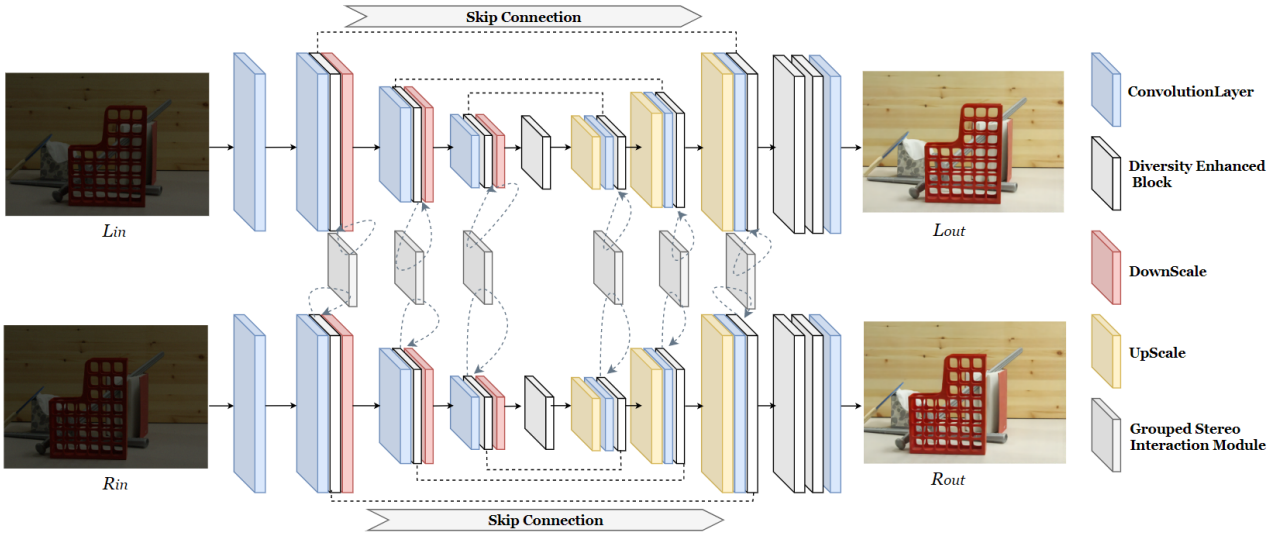

Figure 2: An overview of our proposed DGI-Net. The dotted line in the figure represents the skip connection process, and the dotted arrow represents sending the obtained results to GSIM and returning. L_{in} and R_{in} represents the left and right view images, which are fed into the network simultaneously, and their networks share parameters. The proposed GSIM and DEM focus on completing stereo information interaction, while DEB extracts multi-scale features of feature maps. DEB and GSIM are embedded in DGI-Net, while DEM and group processing strategy are embedded in GSIM.

a solution that exploits Signal-to-Noise-Ratio-aware transformers and convolution network-based modules to dynamically enhance pixels with spatial-varying operations. Zhang et al. [41] proposed DCC-Net to incorporate color information into the LLIE process, thereby obtaining enhanced images with color consistency.

2.2 Stereo Image Restoration

Unlike single image processing, stereo image restoration treats images from left and right views as inputs. Discovering the relationship between stereo pairs and exploiting cross-view cues is crucial for stereo image recovery. For stereo image super-resolution, Jeon et al. [12] proposed a novel method to train two cascaded sub-networks by jointly learning disparity priors, connecting left and right views with different predefined shifts to obtain cross-view information. Ying et al. [37] proposed a parallax attention module (PAM) to capture cross-view cues for stereo image super-resolution. Building on PAM, Wang et al. [31] further incorporated symmetry into cross-view interaction to address practical challenges, such as illumination variation and occlusions. Drawing from NAFNet's success, Chu et al. [2] proposed a simple yet efficient model called NAF-SSR, mainly composed of NAFBlock and a stereo attention module. For stereo image deraining, Zhang et al. [39] proposed a deep network that exploits both stereo images and semantic information to simultaneously solve semantic segmentation and deraining tasks. For stereo image deblurring, Xu and Jia [28] partitioned images into regions according to disparity obtained from stereo blurry images and estimated their blur kernels hierarchically. For low-light stereo image enhancement, Jung et al. [15] proposed a multi-frame GAN to transfer bright scene features from stereo views into poorly lit feature sequences, marking an early attempt at LLSIE. Huang et al. [8] proposed a Laplacian pyramid with the lowest resolution features for view aggregation, effectively saving memory space and enabling feature interaction between stereo pairs. They also designed a wavelet-based view transfer module for efficient multi-scale detail recovery and an illumination-aware attention fusion module to exploit the complementarity between fused features of both views and single-view features. This work has played a significant role in promoting further LLSIE research.

3 PROPOSED METHOD

We first introduce the overall framework of our DGI-Net, and then introduce the designed new modules in detail.

3.1 Overall Framework

The overall architecture of the proposed DGI-Net is shown in Figure 2. As depicted, DGI-Net takes low-light stereo pairs as input and enhances the illumination of images of both left and right views. Note that the weights of dual branches for processing left and right views are shared. Specifically, there are four parts: head, tail, intra-view feature extraction, and inter-view information interaction.

Head and tail. Initially, we use a 3×3 convolution layer to map input stereo images together to feature space for shallow feature extraction. In contrast, we employ two DEBs and one 3×3 convolution layer in the tail to reconstruct the enhanced stereo pairs.

Intra-view feature extraction. To obtain better feature representation, we build a U-shape network with the designed DEB. DEB takes stereo features as input and improves the diversity by utilizing multiple convolution networks with different kernel sizes.

Inter-view information interaction. To facilitate cross-view feature interaction, we insert GSIM into two branches. To be specific, GSIM is capable of capturing cross-view relationship with reduced computational cost and making fuller use of cross-view cues.

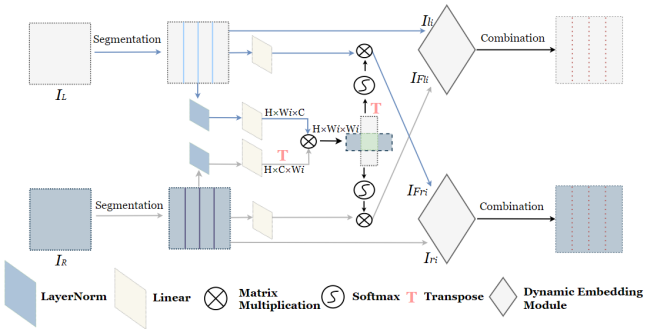

Figure 3: The detailed structure of GSIM. "Segmentation" represents dividing the picture into multiple groups according to the column, and "combination" represents recombining the groups. In order to distinguish easily, the processing progress of the left and right views is indicated by arrows of two colors. We adopt grouping mode to interact dual views and add a DEM to build the dynamic connections between the cross-view cues and intra-view features.

3.2 Grouped Stereo Interaction Module (GSIM)

Since there are much useful information that can be referred to between stereo pairs, it is essential to interact with the inter-view features. Previous works [2, 31, 8] use the matrix multiplication on the entire feature maps to learn the cues between left and right views. However, the computational complexity is quadratic, which implies needing high computational cost when processing high-resolution low-light stereo images. Therefore, we develop GSIM. Different from NAFSSR[2], which introduces high computational cost in image interaction, we incorporates a grouping strategy, as shown in Figure 3. Specifically, we first group the input stereo feature maps with a certain width m as a unit, which can be formulated by

$$I_{l1}, I_{l2}, ..., I_{lm} = Seg(I_l), \quad (1)$$
$$I_{r1}, I_{r2}, ..., I_{rm} = Seg(I_r), \quad (2)$$

where I_l and I_r represent the input stereo feature maps, $Seg(\cdot)$ stands for the grouping operation, m is the number of grouped units of a single view, I_{li} $(i \in [1, m])$ and $I_{ri}(i \in [1, m])$ represent the grouped units of left and right views, respectively. In this way, when performing the matrix multiplication, the calculation amount becomes $1/m^2$ compared to the original one. Further, we send the grouped units into the stereo interaction module. This process can be expressed as follows:

$$I_{Fl_i}, I_{Fr_i} = SIM(I_{li}, I_{ri}), i \in [1, m], \quad (3)$$

where I_{Fl_i} and I_{Fr_i} are the i-th processed units of left and right views, $SIM(\cdot)$ denotes the transformation of cross-view information interaction on two grouped units. Finally, we merge these units to obtain the interacted features as

$$I_{Fl} = Cat([I_{Fl_1}, I_{Fl_2}, ..., I_{Fl_m}]), \quad (4)$$
$$I_{Fr} = Cat([I_{Fr_1}, I_{Fr_2}, ..., I_{Fr_m}]), \quad (5)$$

where $Cat([\cdot])$ denotes the concatenate operation, I_{Fl} and I_{Fr} represent the merged features of left and right views. In this way, the proposed method achieves two purposes: i) enabling cross-view information interaction, ii) reducing the computational complexity from quadratic to linear.

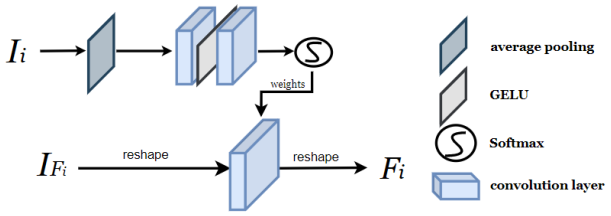

Figure 4: The detailed structure of DEM. I_{Fi} represents the cross-view cues, and I_i represents the intra-view features obtained by directly segmenting the input map, $i \in$[1, n], and F_i represents the stereo interacted feature map after dynamical fusion. DEM incorporates weight learning that aims at building the dynamic connections between the stereo cues and intra-view features.

3.3 Dynamic Embedding Module (DEM)

After discovering the relationship between the left and right views, the next step is using the captured cross-view cues to refine the intra-view features. However, previous works utilize information fusion strategies [8, 31, 15, 2] such as element-wise addition and vanilla convolution-based modules. We argue that this indiscriminate fusion introduces redundant noise while ignoring important cross-view features. This is because the results of cross-view feature fusion are not always beneficial, and some results may introduce substantial noise instead. We should make appropriate choices for this result to promote the restoration of original image details. Different from existing methods, we develop a novel dynamic information fusion strategy termed DEM inspired by dynamic convolution. The detailed structure of DEM is shown in Figure 4. Specifically, we follow previous works using a convolution neural network to fuse the cross-view cues and intra-view features. To build dynamic connections between the cross-view cues and intra-view features, we take two steps. First, we let the weights of the convolution layer be adaptively learned from intra-view features. Second, the learned weights are used to complete convolution operations with the captured cross-view cues. The process can be described as

$$F_i = DWL(I_i) * I_{Fi}, \quad (6)$$

where I_i, I_{Fi} and F_i denote intra-view features, cross-view cues and stereo interacted feature map, respectively; $*$ represents convolution operation, and $DWL(\cdot)$ denotes the transformation of dynamic weights learning, which is composed of two cascaded convolution layers, one average pooling one gelu layer, and one softmax layer. In this way, DEM is capable of achieving dynamic information, which is conducive to promoting the use of cross-view cues.

3.4 Diversity Enhanced Block (DEB)

Current studies paid less attention to intra-view processing and neglected to improve the diversity of extracted features, which might negatively affect the quality of the enhanced results. Simultaneously, they still cannot solve the problem of low contrast and high noise during image enhancement. Therefore, we designed a DEB, which is shown in Figure 6 . The purpose of DEB is to extract the multi-scale features to make the intra-view features stronger and contain more diverse information, and at the same mine potential detail features in low-light images to better guide image detail

Figure 5: Visual comparison results of our DGI-Net with Zero-DCE, Zero-DCE++ and SNR on Flickr1024 dataset. From 2nd column to the last column, the pictures in the 1st row represent reconstructed images, and counterparts in the 2nd row refer to the error maps compared to the ground truth. Less visible content in the error map means a better result. It can be seen from the quantization indicators that our DGI-Net is superior to others. According to the shown error maps, our DGI-Net obtains smallest error which means better detail recovery.

restoration and reduce noise during enhancement. Specifically, we first normalize the input features, and then use a Multilayer Perceptron (MLP) layer to adjust the dimensions of input features and send them into the depth-wise separable convolution neural network (DW-Conv). Note that we use multiple DW-Convs with different kernel sizes. In this way, our DEB is capable of extracting multi-scale features, enhancing the diversity of the captured intra-view features, and further obtaining stronger feature representation. Next, the extracted multi-scale features are concatenated, and two cascaded MLPs are used to fuse these multi-scale features and reduce the channel number, respectively. Residual learning is incorporated for information preservation. The operations in the above process can be expressed as follows:

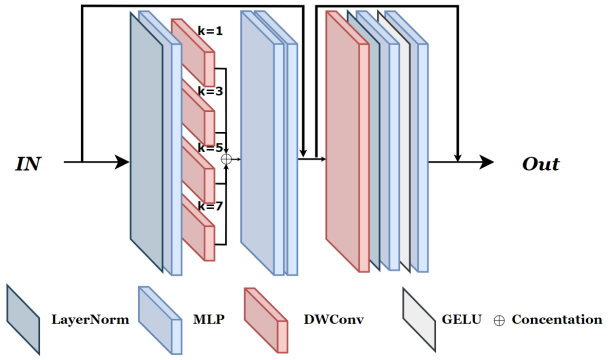

Figure 6: The detailed structure of DEB that is capable of enhancing the diversity of intra-view features via using multiple convolution neural networks with different kernel sizes.

$$
\begin{aligned}
I_m &= \mathrm{MLP}(\mathrm{LN}(I_i)), \\
I_m &= \mathrm{MLP}(\mathrm{MEF}(I_m)), \\
I_m &= I_m + I_i,
\end{aligned} \tag{7}
$$

where I_i and I_m denote the input features and the intermediate features, respectively; $\mathrm{LN}(\cdot)$ and $\mathrm{MLP}(\cdot)$ denote the transformation of layer normalization (LN) and MLP, $\mathrm{MEF}(\cdot)$ represents the transformation of multi-scale feature extraction and fusion. We further utilize a feed-forward network for subsequent processing. As can be seen, there are several components including one DW-Conv layer, one LN layer, two MLPs, and one GELU activation. The detailed process of the feed-forward network can be formulated by:

$$
\begin{aligned}
I_o &= \mathrm{LN}(\mathrm{DW}(I_m)), \\
I_o &= \mathrm{GELU}(\mathrm{MLP}(I_o)), \\
I_o &= \mathrm{MLP}(I_o)+I_m,
\end{aligned} \tag{8}
$$

where I_o denotes the output of DEB, $\mathrm{DW}(\cdot)$ and $GELU(\cdot)$ denote the transformation of depth-wise separable convolution neural network and GELU activation. After this, we can obtain a feature map with noise removed and sufficient feature extraction.

3.5 Loss Functions

To effectively train our DGI-Net, we employ a combination of loss functions that cater to different aspects of the image enhancement process. Specifically, we use the L_1 loss, total variation (TV) loss, and a Fourier-based reconstruction loss. The L_1 loss calculates the difference between the reconstructed image and ground truth, promoting accurate pixel-level recovery:

$$
\mathcal{L}_{cri} = \left\| I_L - \hat{I}_L \right\|_1 + \left\| I_R - \hat{I}_R \right\|_1, \tag{9}
$$

where I_L represents the predicted image, $\hat{I}_L$ represents the ground truth, and $\|\cdot\|_1$ refers to L_1 norm.

Additionally, we use to the total variation loss(TV loss) to better restore the smooth results. The TV loss encourages smoothness in the enhanced images by penalizing abrupt intensity changes, which helps reduce noise and artifacts:

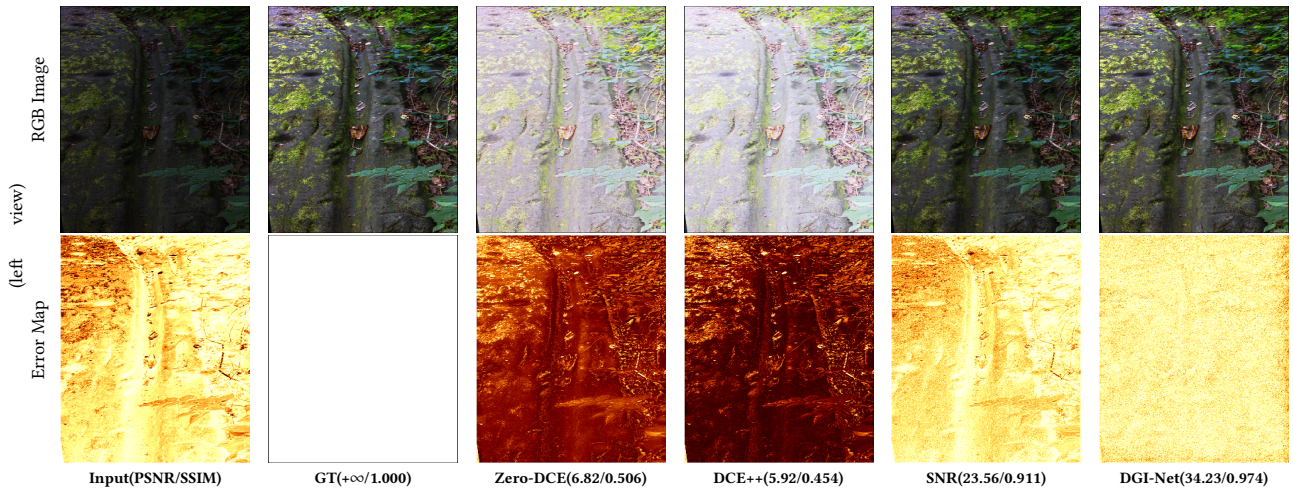

Figure 7: Visual comparison results of our DGI-Net with Zero-DCE, Zero-DCE++ and SNR on Flickr1024 dataset. In each group of results, the 1st row represent reconstructed image, and counterparts in the 2nd row refer to the error maps compared to the ground truth. Less visible content in the error map means a better enhancement result. As can be seen, our proposed DGI-Net better preserves color consistency and restores image details.

Table 1: Numerical evaluation results on Middlebury, Flickr1024, KITTI2012 and KITTI2015 datasets. We show the results on both the left and right views. For better comparison, we report the running time and computational complexity of our model compared to other LLSIE methods. The red font in the table represents the best, and the blue font represents the second.

Methods	GFLOPs/Param	Left(PSNR/SSIM)				Right(PSNR/SSIM)			
		Middlebury	Flickr1024	KITTI2012	KITTI2015	Middlebury	Flickr1024	KITTI2012	KITTI2015
Zero-DCE	-/-	11.34/0.591	10.49/0.420	9.21/0.403	10.17/0.441	11.27/0.592	10.49/0.420	9.18/0.399	10.17/0.438
Zero-DCE++	-/-	10.28/0.563	10.74/0.418	9.15/0.388	10.08/0.427	10.19/0.564	10.74/0.418	9.14/0.384	10.07/0.424
SNR	-/-	21.43/0.851	21.34/0.749	22.11/0.794	20.93/0.769	21.87/0.860	21.46/0.752	22.22/0.793	20.88/0.771
iPASSRNet	23.8/1.42M	21.50/0.872	22.27/0.777	23.32/0.820	21.76/0.802	21.84/0.871	22.13/0.777	23.65/0.829	21.62/0.806
DVENet	2.62/1.17M	20.75/0.855	21.82/0.751	22.14/0.787	21.05/0.767	21.42/0.858	21.50/0.751	21.79/0.787	21.04/0.772
NAFSSR	351.94/23.73M	22.25/0.857	21.68/0.734	21.43/0.775	20.68/0.767	22.54/0.859	21.79/0.737	21.98/0.782	20.60/0.776
DGI-Net(Ours)	4.09/3.22M	25.95/0.932	25.39/0.835	26.86/0.886	24.41/0.836	27.25/0.938	25.06/0.836	26.69/0.890	24.56/0.843

$$\mathcal{L}_{tv} = \sum_{(i,j)\in P,(i,j+1)\in P} \left\| I_{rec}^{i,j+1} - I_{rec}^{i,j} \right\| + \sum_{(i,j)\in P,(i,j+1)\in P} \left\| I_{rec}^{i+1,j} - I_{rec}^{i,j} \right\|, \tag{10}$$

where $\mathcal{L}_{tv}$ denotes the TV loss and I_{rec} stands for the reconstructed image. We employ $\mathcal{L}_{tv}$ on the enhanced images of both views.

The Fourier-based reconstruction loss is designed to enhance high-frequency components of the image, which are crucial for reconstructing sharp edges and fine details in normal-light image:

$$\mathcal{L}_{fre} = \|FFT(I_L) - FFT(\hat{I}_L)\|_1 + \|FFT(I_R) - FFT(\hat{I}_R)\|_1, \tag{11}$$

where $FFT(\cdot)$ stands for the transformation of the fast Fourier transform. By combining these three loss functions, we aim to achieve a balance between accurate pixel-level recovery, smoothness, and preservation of high-frequency details in the enhanced images:

$$\mathcal{L}_{total} = \mathcal{L}_{cri} + \lambda_s \mathcal{L}_{tv} + \lambda_f \mathcal{L}_{fre}, \tag{12}$$

where $\mathcal{L}_{total}$ denotes total loss, λ_s and λ_f denote two hyper-parameters, which are set to 0.1 in this paper.

4 EXPERIMENTS

We evaluate the performance of our DGI-Net on several datasets and show comparison results with other related deep LLIE methods.

4.1 Experimental Settings

Training details. We use the PyTorch framework and train on an NVIDIA GeForce RTX 3090ti GPU with 24GB memory. We use Adam as the optimizer, with an initial learning rate set to 0.0002, which is decayed by a factor of 0.5 every 500 epochs. The batch size is set to 32, and we train for 2000 epochs to ensure convergence.

Evaluated datasets. We test DGI-Net on Middlebury [25] , Flickr1024 [29], KITTI2012 [4] and KITTI2015 [23] datasets. We use 60 stereo paired images from Middlebury and 800 stereo paired images from Flickr1024 as training datasets. For testing, we select 112 paired images from Flickr1024, 20 paired images from KITTI2015, 20 paired images from KITTI2012, and 5 paired images from Middlebury. We follow [38] to synthesize the low-light stereo images and adhere to iPASSRNet's settings [31].

Evaluation metrics. To evaluate the performance of different methods, we choose peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) as evaluation metrics, with larger PSNR and SSIM values indicating better image quality. We also report the number of parameters to test model complexity and evaluate

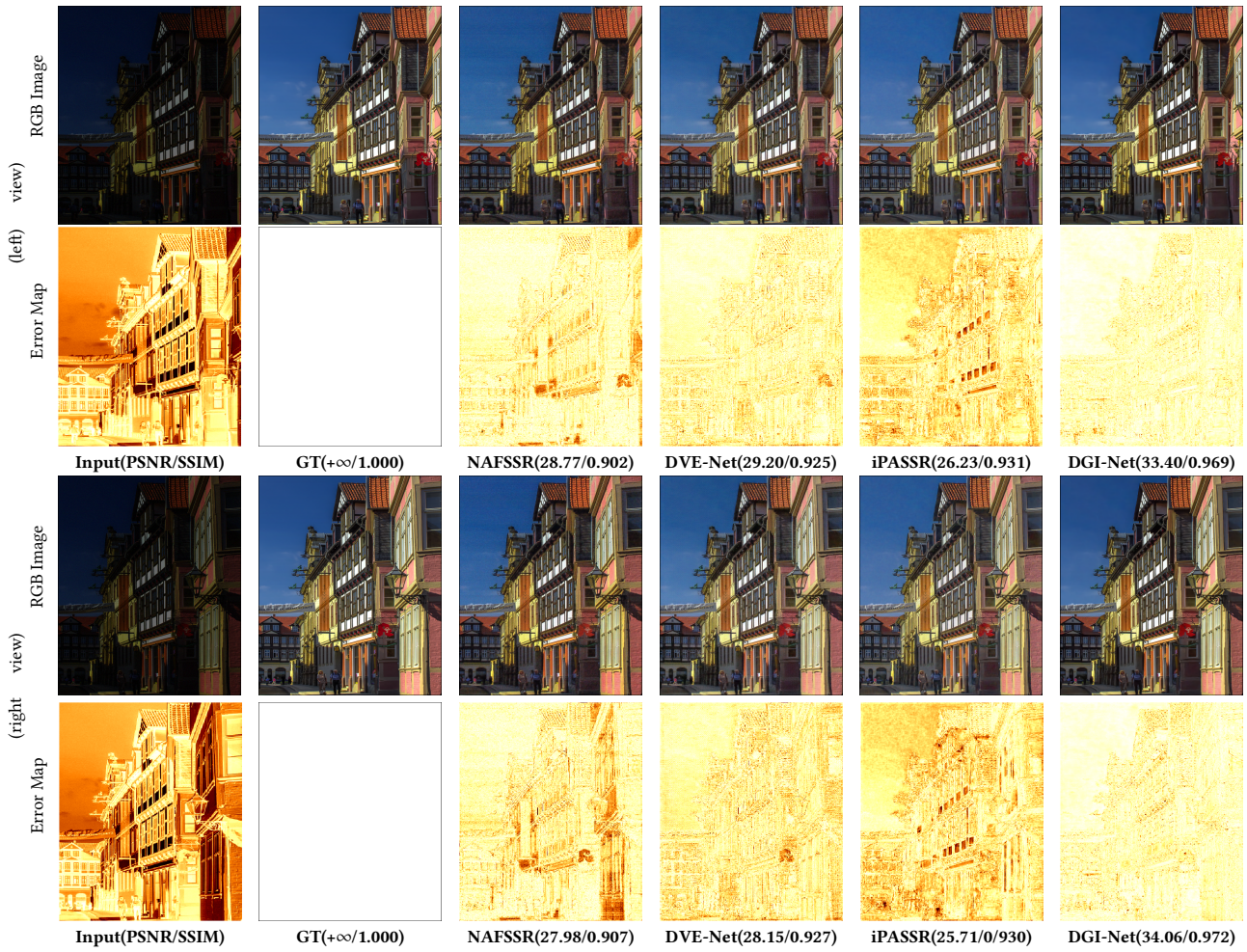

Figure 8: Visual comparison results of our DGI-Net with NAFSSR, DVE-Net and iPASSRNet on Flickr1024 dataset. Specifically, The first group of images is the left view, and the second is the right view. From the 2nd column to the last column, the pictures in the 1st row represent reconstructed images, and counterparts in the 2nd row refer to the error map compared to the ground truth. Less visible content detail in the error map means a better result. From the visible results, we can see that our generated images lose less content detail than other methods.

the computational cost by measuring GFLOPs using torchstat for different models on a 128x128 patch.

Compared methods. We compare DGI-Net with recent representative methods, including single LLIE methods such as Zero-DCE(cvpr'20)[6], Zero-DCE++(tpami'21)[19], and SNR(cvpr'22)[35], as well as LLSIE methods like iPASSRNet (cvprw'21) [31], DVENet (tmm'22)[8], and NAFSSR(cvprw'22)[2]. Since there are currently no more competing models with open-source code dedicated to LLSIE, we include SIR's model for comparison.

4.2 Quantitative Enhancement Results

We evaluate our DGI-Net on Middlebury, KITTI2012, Flickr1024, and KITTI2015 datasets, with numerical results shown in Table 1. We can see that (1) DGI-Net achieves the highest PSNR and SSIM metrics, significantly outperforming exis ting methods, and produces results closest to ground truth. (2) In our measurements, iPASSRNet performs better than other methods. (3) DVENet, a model specifically designed for LLSIE, does not perform exceptionally well. (4) Zero-DCE and Zero-DCE++ yield the worst metrics due to their lack of consideration for correlation features between stereo images. (5) Due to our novel GSIM, our model's computational cost is only slightly higher than that of DVE-Net, while the image enhancement result has been significantly improved.

4.3 Visual Image Analysis results

To provide a more intuitive comparison of performance with other methods, we visualize the results of our main comparison methods on the Middlebury and Flickr1024 datasets. Figures 5 and 7 show our visualization results and corresponding error maps on the Flickr1024 dataset, illustrating the effects of representative single LLIE methods. The process of generating error maps is referenced from [45]. These figures demonstrate that compared to other methods, DGI-Net's enhanced results better restore image details, and the recovered images are closer to ground truth. Moreover, DGI-Net produces lighter error maps compared to other methods. Figures 8 and 1 show the comparison results of our method with other representative LLSIE methods on the Middlebury and Flickr1024 datasets. DGI-Net effectively integrates stereo information while

Table 2: LLSIE results of our DGI-Net with different structures on the Flickr1024 dataset, where the red font in the table represents the best. "normal" in the table represents the use of one 3×3 convolution layer instead of DEB.

DEB	GSIM	DEM	left		right	
			PSNR	SSIM	PSNR	SSIM
×(Normal)	×	×	21.77	0.804	21.69	0.805
✓	×	×	22.69	0.821	22.64	0.821
×(Normal)	✓	×	22.43	0.833	22.43	0.834
×(Normal)	✓	✓	24.83	0.813	24.69	0.814
✓	✓	×	22.38	0.814	22.49	0.816
✓	✓	✓	25.39	0.835	25.06	0.836

introducing less noise, and the reconstructed normal light images are closer to natural lighting. Similar results can be observed from error maps. Overall, our DGI-Net achieves significant performance improvements compared to other methods.

4.4 Ablation Study

Effectiveness of GSIM. To demonstrate the effectiveness of the GSIM module, we test results of removing it from DGI-Net on the Flickr1024 dataset. Table 2 shows the LLSIE results of different models. We find that the effect of GSIM is not so obvious when the DEM module is not added. Comparing the model with only the GSIM module and the baseline model, the performance has improved significantly. This is because GSIM allows sufficient information fusion between stereo images. However, since the model cannot adequately screen the fusion information, it will inevitably introduce redundant features, which will affect the restoration of image details to a certain extent. Only adding GSIM and DEB without adding DEM does not improve the recovery performance. According to the evaluation results, we conclude that visually fused features are not always useful, and adding fused features to the baseline feature map without filtering may interfere with image restoration.

Effectiveness of DEB. From Table 2, we can see that compared to the baseline model, the performance is significantly improved when only adding DEB, which proves the effectiveness of DEB. However, adding DEB with or without screening fusion features can lead to different results: adding DEB while screening fusion features obtains better performance. But when fusion features are not filtered, a large part of the deep layers extracted by DEB is redundant noise from the interaction of stereo information, which will greatly affect image restoration.

Effectiveness of DEM. Since DEM is embedded in the GSIM, we have to add the GSIM in the process of verifying the effect of DEM. From Table 2, we see that adding DEM improves the model's ability to filter effective information for stereo fusion, further facilitating better detail recovery. Besides, DEM also promotes the extraction of effective fusion information by DEB, capturing more details and reducing the influence of noise. In other words, DEM enables DEB and GSIM to work efficiently.

Impact of Different Grouping Strategies. In our experiments, we investigated the impact of the grouping strategy on memory usage and performance by processing images with a size of 256x2048. We cropped and concatenated multiple images on Flickr1024 dataset for testing purposes. Table 3 show that when the width of each group is greater than or equal to 16, the performance remains consistent, with PSNR values around 25.20 and SSIM values around 0.835. However, when the width of each group is less than 16, the performance drops, with PSNR/SSIM values decreasing down to 23.69/0.825 for a group width of 4. In terms of memory usage, we observed that as the group width increases, the maximum memory consumption increases as well. For instance, when the group width is increased from 4 to 1024, the maximum memory usage also increases from 20126.0020 to 30832.1895. It is necessary to note that we were unable to measure the memory consumption when the group width is 2048 due to device limitations, since the Max Memory is over 40960MiB. In summary, our findings suggest that a suitable grouping strategy can maintain performance while managing memory usage, with group widths greater than or equal to 16 yielding the best results.

Table 3: Impact of different group widths on memory usage and performance.

Group Width	Max Memory(MiB)	PSNR	SSIM
4	20126.0020	23.69	0.825
8	20165.7207	24.84	0.831
16	20248.8457	25.23	0.835
32	20416.6582	25.23	0.835
64	20752.1895	25.22	0.834
256	22768.1895	25.19	0.835
1024	30832.1895	25.22	0.835

5 CONCLUSION

We addressed the challenges of dual-view interaction and single-view feature extraction in low-light stereo image enhancement, including high computational cost, inefficient information fusion, and inaccurate feature representation. To overcome these limitations, we proposed a Dual-view Grouped Interaction Network. Specifically, we introduced a Grouped Stereo Interaction Module that incorporates a novel grouping strategy to explore cross-view cues with reduced computational demands. Moreover, we designed a dynamic embedding module to establish dynamic connections between cross-view cues and intra-view features. To enhance feature diversity during intra-view feature extraction, we developed a diversity-enhanced block. Extensive experiments demonstrate the effectiveness and superiority of our DGI-Net for LLSIE.

In future work, we aim to explore the potential application of the interaction mechanism to other models or tasks. As technology advances, developing more efficient and accurate methods for low-light stereo image enhancement is essential to meet the increasing demands of various applications, such as autonomous vehicles and surveillance systems. Our work contributes to this field by introducing an effective and computationally efficient approach, and we hope it inspires future research in low-light stereo image enhancement and related areas.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (62072151, 62106211, 6197607962272142, 61972112), Anhui Provincial Natural Science Fund for the Distinguished Young Scholars (2008085J30), and CAAI-Huawei MindSpore Open Fund. Corresponding author: Zhao Zhang, co-corresponding authors: Yang Zhao and Zhongqiu Zhao.

REFERENCES

[1] Yen-Ching Chang and Chun-Ming Chang. 2010. A simple histogram modification scheme for contrast enhancement. *IEEE Trans. Consumer Electron., 2010.*

[2] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. 2022. NAFSSR: stereo image super-resolution using nafnet. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 2022.*

[3] Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2018. Aesthetic-driven image enhancement by adversarial learning. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 2018.*

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 2012.*

[5] Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, Samuel W. Hasinoff, and Frédo Durand. 2017. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics, 2017.*

[6] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. 2020. Zero-reference deep curve estimation for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 2020.*

[7] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. 2018. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics, 2018.*

[8] Jie Huang, Xueyang Fu, Zeyu Xiao, Feng Zhao, and Zhiwei Xiong. 2022. Low-light stereo image enhancement. *IEEE Transactions on Multimedia, 2022.*

[9] Shih-Chia Huang and Chien-Hui Yeh. 2014. Image contrast enhancement for preserving mean brightness without losing image features. *Eng. Appl. Artif. Intell., 2014.*

[10] John D. Hunter. 2007. Matplotlib: a 2d graphics environment. *Computing in Science Engineering, 2007.*

[11] Yongqing Huo, Fan Yang, Le Dong, and Vincent Brost. 2014. Physiological inverse tone mapping based on retina response. *Vis. Comput., 2014.*

[12] Daniel S. Jeon, Seung-Hwan Baek, Inchang Choi, and Min H. Kim. 2018. Enhancing the spatial resolution of stereo images using a parallax prior. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 2018.*

[13] Haiyang Jiang and Yinqiang Zheng. 2019. Learning to see moving objects in the dark. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019.*

[14] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. 2021. Enlightengan: deep light enhancement without paired supervision. *IEEE Trans. Image Process., 2021.*

[15] Eunah Jung, Nan Yang, and Daniel Cremers. 2019. Multi-frame GAN: image enhancement for stereo visual odometry in low light. In *3rd Annual Conference on Robot Learning, Osaka, Japan, October - November, 2019.* Vol. 100.

[16] Shubhi Kansal, Shikha Purwar, and Rajiv K. Tripathi. 2018. Image contrast enhancement using unsharp masking and histogram equalization. *Multim. Tools Appl., 2018.*

[17] Joung-Youn Kim, Lee-Sup Kim, and Seung-Ho Hwang. 2001. An advanced contrast enhancement using partially overlapped sub-block histogram equalization. *IEEE Trans. Circuits Syst. Video Technol., 2001.*

[18] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. 2022. Low-light image and video enhancement using deep learning: a survey. *Trans. Pattern Anal. Mach. Intell., December 2022.*

[19] Chongyi Li, Chunle Guo, and Chen Change Loy. 2022. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Trans. Pattern Anal. Mach. Intell., 2022.*

[20] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. 2018. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Trans. Image Process., 2018.*

[21] Bin Liu, Weiqi Jin, Yan Chen, Chongliang Liu, and Li Li. 2011. Contrast enhancement using non-overlapped sub-blocks and local histogram projection. *IEEE Trans. Consumer Electron., 2011.*

[22] Artur Loza, David R. Bull, Paul R. Hill, and Alin Achim. 2013. Automatic contrast enhancement of low-light images based on local statistics of wavelet coefficients. *Digit. Signal Process., 2013.*

[23] Moritz Menze and Andreas Geiger. 2015. Object scene flow for autonomous vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 2015.*

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 2015, Proceedings.*

[25] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling. 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition - 36th German Conference, GCPR 2014, Münster, Germany, September 2014, Proceedings.*

[26] Hao Shen, Zhong-Qiu Zhao, Wenrui Liao, Weidong Tian, and De-Shuang Huang. 2022. Joint operation and attention block search for lightweight image restoration. *Pattern Recognition.*

[27] Hao Shen, Zhong-Qiu Zhao, and Wandi Zhang. 2023. Adaptive dynamic filtering network for image denoising. *Proceedings of the AAAI Conference on Artificial Intelligence.*

[28] Bin Sheng, Ping Li, Xiaoxin Fang, Ping Tan, and Enhua Wu. 2020. Depth-aware motion deblurring using loopy belief propagation. *IEEE Trans. Circuits Syst. Video Technol., 2020.*

[29] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jun-Gang Yang, Wei An, and Yulan Guo. 2019. Learning parallax attention for stereo image super-resolution. In *Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 2019.*

[30] Wencheng Wang, Xiaojin Wu, Xiaohui Yuan, and Zairui Gao. 2020. An experiment-based review of low-light image enhancement methods. *IEEE Access, 2020.*

[31] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. 2021. Symmetric parallax attention for stereo image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 2021.*

[32] Yanyan Wei, Zhao Zhang, Yang Wang, Mingliang Xu, Yi Yang, Shuicheng Yan, and Meng Wang. 2021. Deraincyclegan: rain attentive cyclegan for single image deraining and rainmaking. *IEEE Trans. Image Process., 2021.*

[33] Yanyan Wei, Zhao Zhang, Huan Zheng, Richang Hong, Yi Yang, and Meng Wang. 2022. Sginet: toward sufficient interaction between single image deraining and semantic segmentation. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 2022.*

[34] Ruikang Xu, Zeyu Xiao, Mingde Yao, Yueyi Zhang, and Zhiwei Xiong. 2021. Stereo video super-resolution via exploiting view-temporal correlations. In *Proceedings of the ACM International Conference on Multimedia (ACM MM), Virtual Event, China, October 2021.*

[35] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. 2022. Snr-aware low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 2022.*

[36] Bo Yan, Chenxi Ma, Bahetiyaer Bare, Weimin Tan, and Steven C. H. Hoi. 2020. Disparity-aware domain adaptation in stereo image restoration. In *Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, June 2020.*

[37] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Wei An, and Yulan Guo. 2020. A stereo attention module for stereo image super-resolution. *IEEE Signal Process. Lett., Aug 2020.*

[38] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. 2021. Learning temporal consistency for low light video enhancement from single images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 2021.*

[39] Kaihao Zhang, Wenhan Luo, Wenqi Ren, Jingwen Wang, Fang Zhao, Lin Ma, and Hongdong Li. 2020. Beyond monocular deraining: stereo image deraining via semantic understanding. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 2020, Proceedings.*

[40] Zhao Zhang, Yanyan Wei, Haijun Zhang, Yi Yang, Shuicheng Yan, and Meng Wang. 2023. Data-driven single image deraining: a comprehensive review and new perspectives. *Pattern Recognition.*

[41] Zhao Zhang, Huan Zheng, Richang Hong, Mingliang Xu, Shuicheng Yan, and Meng Wang. 2022. Deep color consistent network for low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 2022.*

[42] Suiyi Zhao, Zhao Zhang, Richang Hong, Mingliang Xu, Yi Yang, and Meng Wang. 2022. FCL-GAN: A lightweight and real-time baseline for unsupervised blind image deblurring. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 2022.*

[43] Suiyi Zhao, Zhao Zhang, Richang Hong, Mingliang Xu, Haijun Zhang, Meng Wang, and Shuicheng Yan. 2022. Crnet: unsupervised color retention network for blind motion deblurring. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022.* ACM.

[44] Yang Zhao, Yanbo Ma, Yuan Chen, Wei Jia, Ronggang Wang, and Xiaoping Liu. 2022. Multiframe joint enhancement for early interlaced videos. *IEEE Trans. Image Process.*

[45] Chuanjun Zheng, Daming Shi, and Yukun Liu. 2021. Windowing decomposition convolutional neural network for image enhancement. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 2021.*