This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# Structure Aggregation for Cross-Spectral Stereo Image Guided Denoising

Zehua Sheng[1], Zhu Yu[1], Xiongwei Liu[1], Si-Yuan Cao[1], Yuqi Liu[1], Hui-Liang Shen[1]*, Huaqi Zhang[2]
[1]Zhejiang University, [2]vivo Mobile Communication Company Ltd.
{shengzehua, yu_zhu, liuxw11}@zju.edu.cn karlcao@hotmail.com
{liuyuqi, shenhl}@zju.edu.cn zhanghuaqi@vivo.com

## Abstract

*To obtain clean images with salient structures from noisy observations, a growing trend in current denoising studies is to seek the help of additional guidance images with high signal-to-noise ratios, which are often acquired in different spectral bands such as near infrared. Although previous guided denoising methods basically require the input images to be well-aligned, a more common way to capture the paired noisy target and guidance images is to exploit a stereo camera system. However, current studies on cross-spectral stereo matching cannot fully guarantee the pixel-level registration accuracy, and rarely consider the case of noise contamination. In this work, for the first time, we propose a guided denoising framework for cross-spectral stereo images. Instead of aligning the input images via conventional stereo matching, we aggregate structures from the guidance image to estimate a clean structure map for the noisy target image, which is then used to regress the final denoising result with a spatially variant linear representation model. Based on this, we design a neural network, called as SANet, to complete the entire guided denoising process. Experimental results show that, our SANet can effectively transfer structures from an unaligned guidance image to the restoration result, and outperforms state-of-the-art denoisers on various stereo image datasets. Besides, our structure aggregation strategy also shows its potential to handle other unaligned guided restoration tasks such as super-resolution and deblurring. The source code is available at* https://github.com/lustrouselixir/SANet.

## 1. Introduction

Due to the ill-posed nature of image denoising, even the state-of-the-art single-image denoisers still suffer from the problem of over-smoothing edges and details, especially at high noise levels [5, 26, 35, 41]. Recently, multi-modal image processing has attracted increasing attention in image restoration since it's promising to recover salient structures based on a guidance image [8, 13, 27, 30, 37, 39]. To ensure that the target image keeps the original ambience and the guidance image is nearly noise-free, the auxiliary guidance image is often captured in a different spectral band. In actual applications, near-infrared (NIR) is a popular option [15, 17, 23, 44]. Many smartphones and surveillance systems have already been equipped with NIR cameras. Extra NIR light sources are usually required to capture clean NIR images with salient structures in a short exposure time without affecting the imaging process of the visible light.

Current guided image restoration methods [8, 13, 27, 30, 37] basically assume the target and guidance images to be well-aligned at the pixel level. In [44], the authors apply a beam splitter to obtain registered images from two cameras, while the work [23] uses a motorized rotator with different

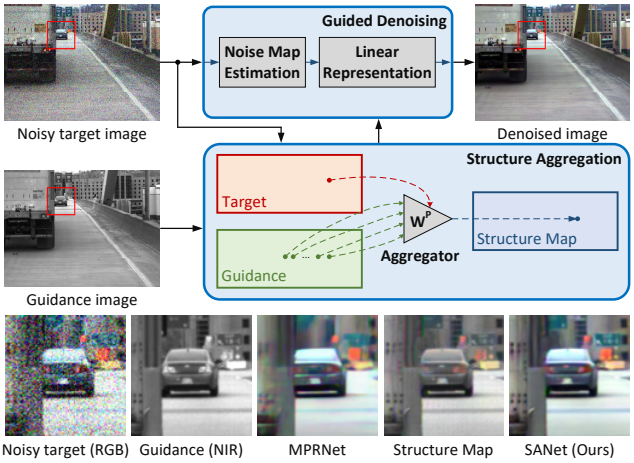


Figure 1. The overall framework of our SANet. It consists of a structure aggregation module and a guided denoising module. The structure aggregation module estimates a structure map for the target image, based on which the guided denoising module regresses the final denoising result. Compared with the state-of-the-art single image denoiser MPRNet [41], our SANet can restore salient structures according to the unaligned guidance image.

*Corresponding author.

filters to separate the visible and NIR information from the mixed signals. However, these special designs are too complex to be directly deployed into portable devices. A more common way to record information from different spectra is to use a stereo camera system, making it essential to conduct guided denoising for unaligned cross-spectral images.

One straightforward solution is to register the input image pairs through stereo matching. However, current cross-modal stereo matching methods rarely consider the case of noise contamination and precise disparity estimation is intractable. Obtaining a large number of ground-truth disparities for supervised training is also a quite heavy work. In addition, as stated in [38], even an accurate pixel-level registration result may be a sub-optimal representation for a vision task due to the occlusion problem. Therefore, to handle guided denoising of unaligned images, we do not consider stereo matching as an essential step.

The guided image filtering theory [13, 27] actually introduces an inspirational idea, that the target image can be regressed by a spatially variant linear representation model of the guidance image as long as they are structurally aligned. The pixel and gradient intensities of the guidance image are not crucial factors affecting the accuracy of guided denoising. Inspired by this, in this work, we propose a guided denoising framework for cross-spectral stereo image pairs. Instead of warping the guidance image based on the one-to-one pixel correspondence constructed by stereo matching, we estimate a structure map for the target image by aggregating non-local information from the guidance image.

Here, we assume that the input paired images have been rectified, so that they only have horizontal disparities. To restore a target pixel, we extract features around pixels within the range of the maximum disparity in the guidance image. Based on the structural correlation between these candidate pixels and the noisy target one, we further fuse them to estimate the structures around this target pixel. The restoration result can then be computed with a spatially variant linear model to adjust the pixel and structural intensities.

Based on the above analysis, we design a convolutional neural network, called as SANet, to handle guided denoising for cross-spectral stereo image pairs. Its overall framework is shown in Fig. 1. A structure aggregation module extracts non-local features from the guidance image and computes their perceptual correlation with the noisy target ones to estimate a clean structure map. Then, a guided denoising module regresses the denoising result from the estimated structure map with a spatially variant linear representation model. In the training stage, we don't require ground-truth disparity information to optimize the network. Due to the redundancy of image structures, the occluded contents can also be properly estimated with high probability according to their adjacent information.

In summary, the main contributions of this work are as follows: (1) For the first time, we propose a guided denoising framework for cross-spectral stereo image pairs, which aims to recover salient structures of the target noisy image from an unaligned guidance image. (2) We introduce a noise-robust structure aggregation strategy to estimate a clean structure map for the noisy target image without conventional stereo matching, based on which we restore the final denoising result. (3) Experimental results show that our algorithm outperforms state-of-the-art denoising methods on various datasets, and our structural aggregation strategy also has the potential to handle other unaligned guided restoration tasks such as super-resolution and deblurring.

## 2. Related Work

### 2.1. Single-Image Denoising

Single-image denoising aims to estimate the latent clean image from one single noisy observation. Traditional algorithms can be categorized into filtering-based ones [3, 7] and optimization-based ones. The optimization-based algorithms model denoising with objective functions regularized by image priors such as sparsity [9, 36] and low-rankness [10, 21], which often require a long inference time.

Currently, the deep learning theory has greatly improved both denoising accuracy and efficiency. DnCNN [42] outperforms traditional Gaussian denoisers using a simple network with batch normalization and residual learning. To handle blind denoising, CBDNet [11] trains a noise estimation sub-network. More recently, MIRNet [40] aggregates contextual information using parallel multi-resolution convolution streams and attention mechanism. MPRNet [41] designs a multi-stage network architecture to progressively learn the restoration function. In [4], HINet shows the contributions of instance normalization in image restoration. Built on the Locally-enhanced Window Transformer block, Uformer [35] introduces a U-shaped Transformer to handle both local context and long-range dependencies more efficiently. However, despite their strong abilities to remove noise, the single-image denoisers still inevitably over-smooth edges and textures, especially at high noise levels.

### 2.2. Guided Image Restoration

Current studies on guided image restoration basically assume that the degraded target image and the guidance image are well-aligned. The pioneering work, guided image filtering [13] shows that the target image can be linearly represented by the guidance image in local windows. The work [39] uses a scale map to accomplish adaptive smoothing and edge preservation at the same time.

The deep learning theory has also been introduced into joint restoration tasks. SVLRM [27] and UMGF [31] restore the target image by learning a spatially variant representation model of the guidance image, while FGDNet [30]

exploits frequency decomposition to well balance noise removal and structure preservation. Representing the common and unique features of images, CUNet [8] can handle both guided restoration and guided fusion tasks. In [37], guided denoising is modeled as as an optimization problem regularized by a deep implicit prior.

However, few studies are conducted under unaligned situations. In [32], the authors perform patch-wise guided filtering using a set of translated guidance images, but it cannot handle stereo image pairs as the misalignment involves not only translation but also rotation. Besides, the iterative optimization process can be computationally expensive in the case of large displacement, and the denoising ability of patch-wise guided filtering is limited. UGSR [12] super-resolves thermal images guided by unaligned RGB images. To our best knowledge, there is still no effective solution for unaligned guided image denoising, especially when the image pairs are captured by stereo cameras.

### 2.3. Cross-Modal Stereo Matching

Traditional cross-modal stereo matching algorithms focus on exploring similarity metrics across different modalities. Based on normalized cross-correlation, RSNCC [29] can deal with structure divergence caused by inconsistent shadows and reflections on the object surfaces. DASC [18] describes image structures according to the self-correlation of patches in a local support window. Further, the work [1] uses neural networks to measure the similarity of cross-modal patches. To refine the depth information recorded by the RGB-D cameras, the work [28] focuses on stereo matching between RGB images and depth maps. In [45], the authors conduct stereo matching with material-aware confidence. Its spectral translation network is specially designed for RGB-NIR stereo images. In comparison, our proposed SANet can deal with more general cross-spectral situations. The work [20] transforms images across different spectral bands through adversarial learning. However, it does not consider the case of noise contamination.

## 3. Method

### 3.1. Model Formulation

Denote $\mathbf{X},\ \mathbf{G} \in \mathbb{R}^{H\times W}$ as a pair of images captured in different spectral bands. $\mathbf{X}$ is our desired target image but corrupted by additive noise $\mathbf{N} \in \mathbb{R}^{H\times W}$. We only have access to its noisy observation $\mathbf{Y} = \mathbf{X} + \mathbf{N}$. $\mathbf{G}$ is the clean guidance image. If they are well-aligned, as stated in [27], the clean target image can be estimated by a spatially variant linear representation of the guidance image, computed by

$$\hat{\mathbf{X}}(i,j) = \mathbf{A}(i,j)\cdot\mathbf{G}(i,j) + \mathbf{B}(i,j). \tag{1}$$

Here, $(i,j)$ are the pixel coordinates, $\mathbf{A},\ \mathbf{B} \in \mathbb{R}^{H\times W}$ are the linear coefficient matrices.

However, if the input paired images are not well-aligned, for instance, captured by a cross-spectral stereo camera system, this linear representation model may lead to image distortions due to their structural inconsistencies. To address this, stereo matching is a direct solution, but it's hard to construct accurate one-to-one correspondences of pixels across different modalities, especially in the presence of noise.

In fact, the guided filtering theory implies that, the structures of the guidance image can be properly transferred to the restoration target as long as they are structurally aligned. That is, if we expect to align the guidance image with the target one, it doesn't necessarily need to retain the original pixel and gradient intensities, nor to satisfy the one-to-one correspondence of pixels constrained by stereo matching.

Motivated by this, in this work, we propose to aggregate structural information from the guidance image to estimate a structure map for the noisy target image. We assume that the noisy target $\mathbf{Y}$ and the guidance $\mathbf{G}$ are captured in the left and the right views, respectively. The stereo image pairs are rectified so they only have horizontal disparities. Denote $D$ as the maximum disparity value. That is, for a pixel $\mathbf{Y}(i,j)$ in the target image, its correspondent pixel in the guidance image is $\mathbf{G}(i-r,j),\ 0 \le r \le D$.

To restore a target pixel, instead of matching only one single pixel in the guidance image, we aggregate structures around all pixels within the range of the maximum disparity. These candidate pixels in the guidance image are then weighted averaged to generate the desired structures that are consistent with the noisy target. Hence, our guided denoising model for stereo images can be formulated as

$$\hat{\mathbf{X}}(i,j) = \sum_{d=0}^{D} \mathbf{W}_d(i,j)\cdot\mathbf{G}(i-d,j) + \mathbf{B}(i,j), \tag{2}$$

where $\mathbf{W}_d$ and $\mathbf{B}$ are the weight matrix and the bias term, respectively. Further, to enable the aggregating process to focus more on structural correspondence than structural intensities, we split $\mathbf{W}$ into two components, *i.e.*, a perceptual weight $\mathbf{W}^P$ and a scale weight $\mathbf{W}^S$. The perceptual weight aims to regress the target structures by aggregating information from candidate pixels, which we also refer to as a structure aggregator. The scale weight is used to adjust the structural intensity to match the target image. Therefore, Eq. (2) can be re-written as

$$\begin{aligned}\hat{\mathbf{X}}(i,j) &= \mathbf{W}^S(i,j)\cdot\sum_{d=0}^{D}\mathbf{W}_d^P(i,j)\mathbf{G}(i-d,j) + \mathbf{B}(i,j)\\ &= \mathbf{W}^S(i,j)\cdot\mathbf{U}(i,j) + \mathbf{B}(i,j)\end{aligned}, \tag{3}$$

where $\mathbf{U}$ is our estimated structure map. In other words, the denoising result can be estimated using a spatially variant linear representation of this structure map.

In the actual implementation, we circularly translate the

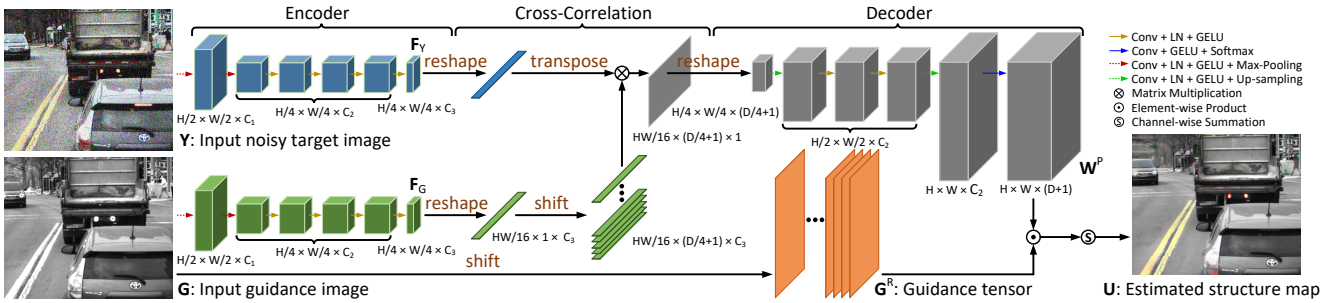


Figure 2. The network architecture of the structure aggregation module. First, two separate encoders extract feature maps from the input noisy target and the guidance images. Then, the feature maps of the guidance image are circularly shifted to compute the feature-wise cross-correlation with the target image. Finally, a decoder is used to generate the perceptual weight from the correlation map to estimate the structural map for the target image.

guidance image to the right for $D$ times. Stacking the original and the translated guidance images together in the channel direction, we can obtain a guidance tensor $\mathbf{G}^R \in \mathbb{R}^{H\times W\times(D+1)}$ with $D+1$ channels. We re-write Eq. (3) as

$$\hat{\mathbf{X}} = \mathbf{W}^S \odot \mathbf{U} + \mathbf{B} = \mathbf{W}^S \odot \mathcal{S}\big(\mathbf{W}^P \odot \mathbf{G}^R\big) + \mathbf{B}, \quad (4)$$

where $\mathbf{W}^S, \mathbf{U}, \mathbf{B} \in \mathbb{R}^{H\times W}$, and $\mathbf{W}^P \in \mathbb{R}^{H\times W\times(D+1)}$. $\mathcal{S}(\cdot)$ is the channel-wise summation function and $\odot$ is the element-wise product operator. Further, to strike a good balance between structure restoration and noise removal, following the work [30], we construct the spatially variant representation model in the frequency domain.

In this work, we design a convolutional neural network, called as SANet, to accomplish the entire denoising process. The overall architecture of our proposed SANet is shown in Fig. 1. It consists of a structure aggregation module and a guided denoising module. We will give a detailed introduction of them in the following subsections.

### 3.2. Structure Aggregation Module

Compared with conventional stereo matching that aims to construct positional correspondences, we focus on learning structural correlations across different modalities to estimate the structure map for guided denoising. Therefore, instead of cost aggregation and disparity estimation, the structure aggregation module predicts a perceptual weight $\mathbf{W}^P$ to fuse the candidate pixels of the guidance image within the range of the maximum disparity based on their feature-wise correlation with the target pixel.

Its architecture is shown in Fig. 2. First, two identical encoders extract two feature maps, $\mathbf{F}_Y$ and $\mathbf{F}_G$, from the input images $\mathbf{Y}$ and $\mathbf{G}$. The encoder contains 6 convolution layers with kernels of size $3\times 3$. Each convolution is followed by layer normalization [2] and GELU [14]. Feature maps produced by the 1st convolution layer have $C_1$ channels. The 2nd$\sim$5th layers produce feature maps with $C_2$ channels, while the 6th one generates $\mathbf{F}_Y$ and $\mathbf{F}_G$ with $C_3$ channels. In this work, we set $C_1 = 48$, $C_2 = 96$, and $C_3 = 24$. The $2\times 2$ max-pooling operation with stride 2 is used in the first two convolution layers for down-sampling.

To evaluate the feature-wise correlation between the target and the guidance images within the range of the maximum disparity, we circularly shift $\mathbf{F}_G$ to the right for $D/4$ times. Stacking them with the original $\mathbf{F}_G$, we obtain a feature tensor of size $H/4\times W/4\times(D/4+1)\times C_3$. It is then used to compute the channel-wise cross-correlation with the transposed $\mathbf{F}_Y$ via matrix multiplication. The size of the correlation map is $H/4\times W/4\times(D/4+1)\times 1$.

Then, the correlation map is fed into a decoder consisting of 5 convolution layers to compute the perceptual weight $\mathbf{W}^P$. Each convolution operation is also followed by layer normalization and GELU. The feature maps computed in the first 4 layers have $C_2$ channels. They are up-sampled in the 1st and the 4th convolution layers by bilinear interpolation. A channel-wise Softmax is then computed over the predicted $\mathbf{W}^P$. Finally, the structure map $\mathbf{U}$ is obtained by computing the element-wise product of $\mathbf{W}^P$ and the guidance tensor $\mathbf{G}^R$ followed by a channel-wise summation.

### 3.3. Guided Denoising Module

The guided denoising module aims to obtain the restoration result $\hat{\mathbf{X}}$ using a spatially variant linear representation of the structure map $\mathbf{U}$. In this work, we follow the network architecture of [30] with slight adjustments to learn the linear representation model in the frequency domain. As displayed in Fig. 1, the guided denoising module contains two steps: noise map estimation and linear representation.

**Noise map estimation** gives an initial noise estimate according to the noisy target image to improve the robustness of guided denoising at different noise levels. It consists of 16 convolution layers. The first 15 layers produce feature maps with 64 channels. Each $3\times 3$ convolution operation is followed by layer normalization and GELU except for the last one that outputs a single-channel noise map $\hat{\mathbf{N}}$.

**Linear representation** aims to predict a scale weight and

a bias term to regress the final denoising result. Here, the scale weight adjusts the structure intensities of the structure map $\mathbf{U}$, while the bias term ensures that the basic pixel intensities of the denoising result are faithfully restored according to the noisy target image. Motivated by [30], we compute the bias term based on the frequency coefficients of the noisy target and the estimated noise map.

First of all, we transform $\mathbf{Y}$, $\hat{\mathbf{N}}$ and $\mathbf{U}$ into the frequency domain using patch-wise 2D discrete cosine transform (2D-DCT) and stack the frequency coefficients into three frequency tensors of size $H \times W \times k^2$, where $k$ is the patch size. The frequency tensors are separately fed into three identical encoders to extract three feature maps, which are then concatenated and fed into three decoders to predict three weight tensors of size $H \times W \times k^2$. The frequency coefficients of the denoised image are computed by weighted averaging the three frequency tensors, which are later transformed back to the spatial domain with the inverse 2D-DCT. Here, the encoders and decoders share the same architectures as those in the structure aggregation module except for the channel numbers of the feature maps. The 1st and the 6th layers of the encoder produce features with 96 channels, while the 2nd$\sim$5th ones produce 128-channel feature maps. Each decoder predicts a $k^2$-channel weight tensor to regress the frequency coefficients of the target image. Feature maps obtained by the first 4 layers of the decoder have 96 channels.

### 3.4. Loss Function

We use a structure aggregation loss $\mathcal{L}_{\text{sa}}$ and a guided denoising loss $\mathcal{L}_{\text{gd}}$ to separately optimize the structure aggregation module and guided denoising module.

**Structure aggregation loss**. For structure aggregation, we adopt the perceptual loss [16] constructed based on the pre-trained VGG-16 network as $\mathcal{L}_{\text{sa}}$ to minimize the structural distance between the estimated structure map $\mathbf{U}$ and the ground-truth restoration target $\mathbf{X}$.

**Guided denoising loss**. The guided denoising loss function $\mathcal{L}_{\text{gd}}$ consists of a noise estimation loss and a linear representation loss to ensure the overall denoising accuracy. We implement them using the smooth L1 losses , formulated as

$$\mathcal{L}_{\text{gd}} = \mathcal{L}_{1,\text{smooth}}\big(\mathbf{X}, (\mathbf{Y} - \hat{\mathbf{N}})\big) + \mathcal{L}_{1,\text{smooth}}\big(\mathbf{X}, \hat{\mathbf{X}}\big). \quad (5)$$

## 4. Experiments

In this section, we evaluate the proposed SANet in two scenarios, including RGB-NIR and synthetic cross-spectral stereo image pairs. During noise removal, the network processes one channel at a time. That is, for color images, their R, G, and B channels are denoised separately. We adopt peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) to assess the denoising accuracy. Learned perceptual image patch similarity (LPIPS) [43] is used to evaluate the visual quality. Higher PSNR and SSIM and lower LPIPS values indicate better performance.

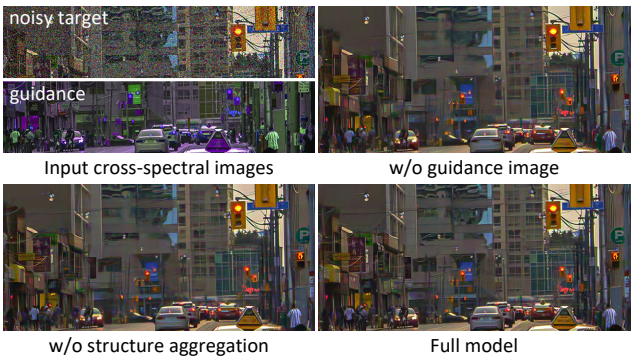


Figure 3. Visual comparison of denoising results obtained by our SANet in three cases: (i) without guidance image, (ii) without structure aggregation, (iii) full model.

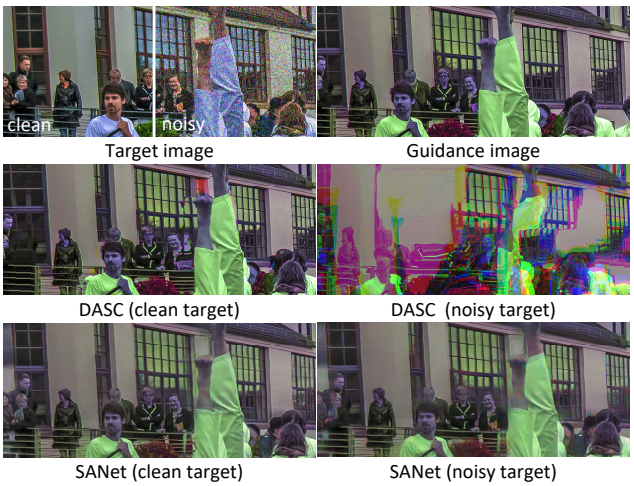


Figure 4. Visual comparison of the warped guidance images obtained by DASC [18] and the structure maps estimated by our SANet in the case of clean target image and noisy target image ($\alpha = 0.02$, $\sigma = 0.2$), respectively.

### 4.1. Experimental Settings

**Datasets**. For the task of NIR-guided RGB image denoising, the algorithms are evaluated on the PittsStereo-RGBNIR Dataset [45] with 40000 paired training data and 2000 paired test data. The stereo image pairs in this set have quite small disparities, and are acquired in similar scenes.

To further evaluate SANet in more challenging scenarios including large disparities and different cross-spectral cases, we also simulate cross-spectral image pairs from the RGB stereo datasets, including the Flickr1024 [34] and the KITTI Stereo 2015 Dataset [25]. The former consists of 800 paired images for training and 112 for evaluation, while the latter contains 400 paired images for training and 400 for

| Methods | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| w/o structure aggregation | 25.30 | 0.8345 | 0.2945 |
| w/o guidance image | 25.26 | 0.8313 | 0.2983 |
| w/o noise estimation | 25.48 | 0.8435 | 0.2690 |
| full model | **25.67** | **0.8477** | **0.2685** |

Table 1. Ablation studies on structure aggregation, guidance image, and noise estimation under Gaussian noise ($\sigma = 0.2$).

| Algorithms | $\alpha = 0, \sigma = 0.2$ | | | $\alpha = 0.02, \sigma = 0.2$ | | |
|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| MIRNet [40] | 27.72 | 0.8391 | 0.4029 | 27.30 | 0.8327 | 0.4137 |
| NBNet [5] | 27.44 | 0.8340 | 0.4066 | 27.00 | 0.8274 | 0.4099 |
| MPRNet [41] | 27.74 | 0.8416 | 0.3861 | 27.34 | 0.8349 | 0.3967 |
| HINet [4] | 27.94 | 0.8448 | 0.3922 | 27.55 | 0.8386 | 0.4029 |
| Uformer [35] | 27.67 | 0.8383 | 0.3985 | 27.26 | 0.8319 | 0.4023 |
| DGUNet [26] | 27.79 | 0.8390 | 0.3825 | 27.39 | 0.8336 | 0.3913 |
| FGDNet [30] | 28.31 | 0.8540 | 0.3218 | 27.94 | 0.8493 | 0.3280 |
| MNNet [37] | 29.12 | 0.8742 | 0.2577 | 28.80 | 0.8699 | 0.2654 |
| SANet (Ours) | **29.32** | **0.8761** | **0.2565** | **28.98** | **0.8726** | **0.2606** |

Table 2. The average PSNR (dB), SSIM, and LPIPS values of single-image [4,5,26,35,41,41] and guided [30,37] denoisers and our SANet on the PittsStereo-RGBNIR Dataset under Gaussian noise ($\sigma = 0.2$) and Poisson-Gaussian noise ($\alpha = 0.02, \sigma = 0.2$).

evaluation. Specifically, we take the right-view images as the guidance, and use their G, B, and R channels to guide the denoising process of the R, G, and B channels of the left-view noisy target images, respectively. Here, the network processes one channel at a time. In this case, we can ensure that the input images are from different spectra. For clear illustration, we change the channel orders of the guidance images to GBR for display in the following figures.

**Training details**. We train the network using the Adam optimizer [19] with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The weight decay is set to $1 \times 10^{-8}$, and the batch size is 8. The training process contains two stages. The structure aggregation module is optimized with the loss function $\mathcal{L}_{sa}$ in the first stage. In the second stage, we only optimize the guided denoising module using $\mathcal{L}_{gd}$. Each optimization process takes $2 \times 10^5$ iterations with the initial learning rate $1 \times 10^{-4}$ gradually reduced to $1 \times 10^{-6}$ using the cosine annealing schedule [22]. In this work, we deal with Poisson noise and Gaussian noise, the two main noise types in modern camera systems. The noise levels are indicated by $\alpha$ and $\sigma$, respectively. In each iteration, we randomly crop patches of size $128 \times 400$ from the source paired images as training data. The noisy target patches are simulated with random $\alpha$ ranging from 0 to 0.02 and $\sigma$ ranging from 0 to 0.2. For synthetic cross-spectral cases, the target and the guidance images are randomly extracted from two different channels of the RGB stereo images in the training stage. The maximum disparity $D$ is set to 128. The patch size of 2D-DCT in guided denoising is set to $k = 9$. All experiments are conducted using one Nvidia Quadro RTX8000 GPU.

| Algorithms | $\alpha = 0, \sigma = 0.2$ | | | $\alpha = 0.02, \sigma = 0.2$ | | |
|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| MIRNet [40] | 25.26 | 0.8324 | 0.3061 | 24.88 | 0.8244 | 0.3158 |
| NBNet [5] | 24.86 | 0.8216 | 0.3207 | 24.47 | 0.8141 | 0.3258 |
| MPRNet [41] | 25.33 | 0.8370 | 0.2968 | 24.95 | 0.8287 | 0.3058 |
| HINet [4] | 25.53 | 0.8426 | 0.2941 | 25.16 | 0.8347 | 0.3020 |
| Uformer [35] | 25.15 | 0.8302 | 0.3052 | 24.76 | 0.8225 | 0.3102 |
| DGUNet [26] | 25.32 | 0.8294 | 0.2953 | 24.96 | 0.8229 | 0.3043 |
| FGDNet [30] | 24.97 | 0.8185 | 0.3107 | 24.54 | 0.8102 | 0.3189 |
| MNNet [37] | 25.12 | 0.8269 | 0.3181 | 24.73 | 0.8173 | 0.3259 |
| SANet (Ours) | **25.67** | **0.8477** | **0.2685** | **25.30** | **0.8411** | **0.2736** |

Table 3. The average PSNR (dB), SSIM, and LPIPS values of single-image [4,5,26,35,41,41] and guided [30,37] denoisers and our SANet on the Flickr1024 Dataset under Gaussian noise ($\sigma = 0.2$) and Poisson-Gaussian noise ($\alpha = 0.02, \sigma = 0.2$).

| Algorithms | $\alpha = 0, \sigma = 0.2$ | | | $\alpha = 0.02, \sigma = 0.2$ | | |
|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| MIRNet [40] | 26.72 | 0.8619 | 0.3276 | 26.30 | 0.8553 | 0.3375 |
| NBNet [5] | 26.52 | 0.8570 | 0.3366 | 26.09 | 0.8501 | 0.3423 |
| MPRNet [41] | 26.59 | 0.8580 | 0.3308 | 26.18 | 0.8509 | 0.3407 |
| HINet [4] | 26.92 | 0.8668 | 0.3131 | 26.52 | 0.8607 | 0.3224 |
| Uformer [35] | 26.68 | 0.8609 | 0.3270 | 26.27 | 0.8546 | 0.3313 |
| DGUNet [26] | 26.55 | 0.8577 | 0.3287 | 26.12 | 0.8505 | 0.3415 |
| FGDNet [30] | 26.77 | 0.8623 | 0.3111 | 26.34 | 0.8559 | 0.3198 |
| MNNet [37] | 27.33 | 0.8742 | 0.2884 | 26.89 | 0.8680 | 0.2973 |
| SANet (Ours) | **27.88** | **0.8899** | **0.2439** | **27.47** | **0.8851** | **0.2513** |

Table 4. The average PSNR (dB), SSIM, and LPIPS values of single-image [4, 5, 26, 35, 41, 41] and guided [30, 37] denoisers and our SANet on the KITTI Stereo 2015 Dataset under Gaussian noise ($\sigma = 0.2$) and Poisson-Gaussian noise ($\alpha = 0.02, \sigma = 0.2$).

## 4.2. Ablation Studies

In this section, we validate the influence of guidance image, structure aggregation, and noise estimation on the denoising performance of our algorithm on the Flickr1024 Dataset under Gaussian noise with $\sigma = 0.2$. In the case of evaluating our SANet without guidance image, for a fair comparison, we use the noisy target image to guide its own denoising process to retain the original network architecture. As shown in Table 1, the denoising accuracy decreases when any of the three components is deactivated. Without guidance image, SANet is equivalent to a single-image denoiser and thus cannot preserve fine structures during noise removal. Without structure aggregation, the structures of the guidance image cannot be properly transferred to the denoising result. A visual comparison of the ablation results

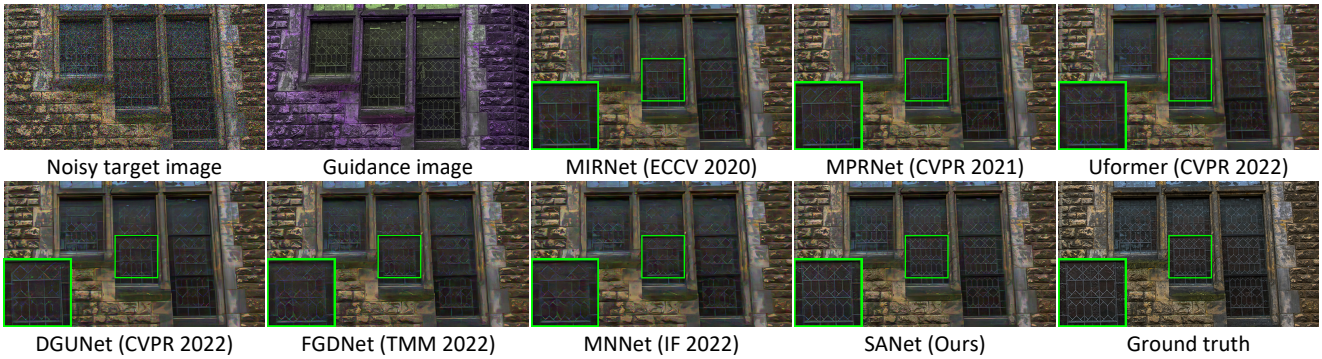


Figure 5. Denoising results on the Flickr1024 Dataset under Gaussian noise ($\sigma = 0.2$) obtained by the comparative denoising methods and our SANet.

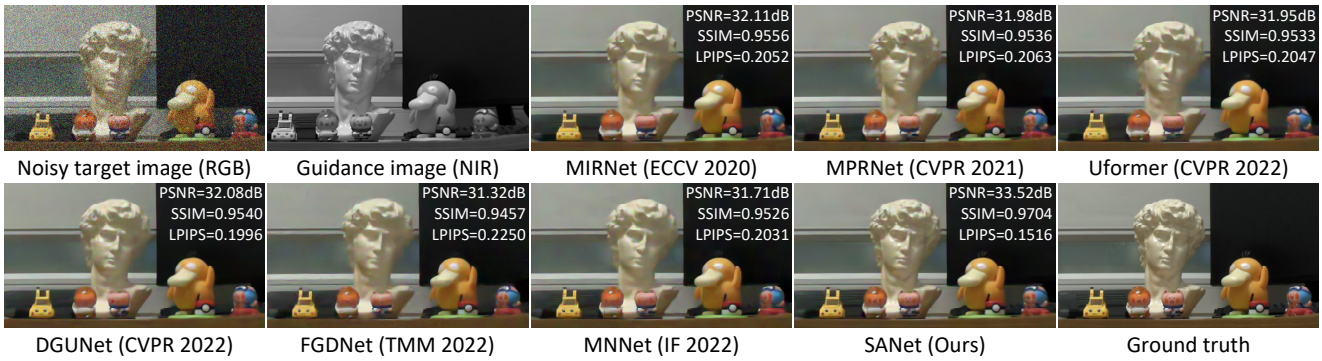


Figure 6. Denoising results of our captured RGB-NIR stereo image pairs under Gaussian noise ($\sigma = 0.2$) obtained by the comparative denoising methods and our SANet.

is displayed in Fig. 3.

### 4.3. Evaluation and Comparison

To show the robustness of our structure aggregation strategy in the case of noise contamination, we compare our estimated structure map to the warped guidance image obtained by stereo matching using DASC [18]. As shown in Fig. 4, except in those occluded regions, DASC can achieve plausible matching result when the target image is clean. However, its matching accuracy is severely decreased in the presence of noise, based on which it's difficult to transfer structures to the final restoration target. More detailed results and the evaluation of other cross-spectral stereo matching methods are displayed in the supplementary materials.

In comparison, our proposed structure aggregation process can estimate an accurate structure map for the target image whether it is corrupted by noise or not. When there occur structural inconsistencies such as occlusion, it can predict the occluded structures as much as possible based on the adjacent information. As shown in Fig. 4, even the missing contents on the left side of the guidance image can also be properly completed after structure aggregation.

Further, we quantitatively evaluate our SANet and compare it to the state-of-the-art single-image denoisers including MIRNet [40], NBNet [5], MPRNet [41], HINet [4], Uformer [35], and DGUNet [26]. To show the superiority of SANet in handling unaligned situations, we also compare it with two latest guided denoisers FGDNet [30] and MNNet [37], which are designed for aligned situations. All competing methods are re-trained using the same training set as ours for fair comparisons. For single-image denoisers, we also use an additional dataset of 3859 images [33] to optimize the networks. Each channel of the color image is processed separately for both comparative and our models.

Tab. 2 lists the average PSNR, SSIM and LPIPS values on the RGB-NIR image pairs from the PittsStereo-RGBNIR test dataset. Our SANet achieves the highest denoising accuracy. Since image pairs from this dataset have quite small disparities, image structures are likely to remain consistent in some regions due to their continuity and redundancy natures in local areas. Therefore, the comparative guided denoising models FGDNet and MNNet still achieve better denoising results than the single-image ones. Tab. 3 and Tab. 4 list the quantitative results on the synthetic cross-spectral data generated from the Flickr1024 and the KITTI Stereo 2015 Datasets, where the disparities of the input im-

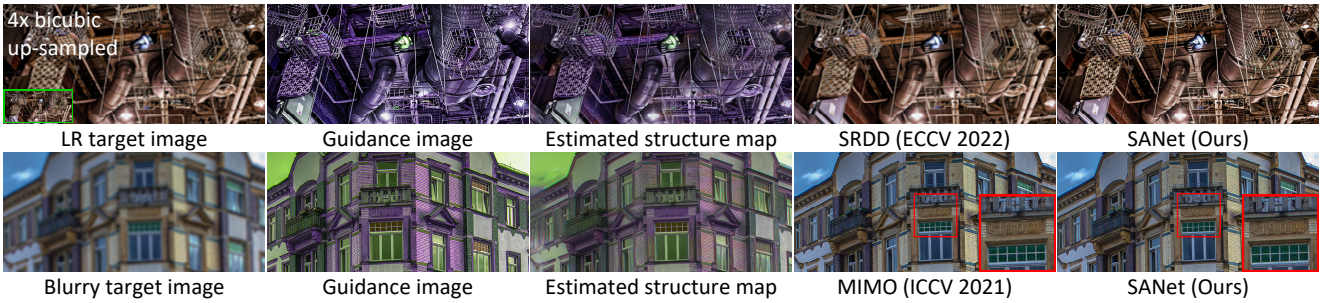


Figure 7. Guided super-resolution (top row) and guided deblurring (bottom row) results of synthetic cross-spectral stereo image pairs on the Flickr1024 Dataset. The low-resolution (LR) target image is generated with 4× bicubic down-sampling. The blurry target image is generated with a Gaussian kernel of standard deviation 5.

| | MIRNet [40] | NBNet [5] | MPRNet [41] | HINet [4] | Uformer [35] | DGUNet [26] | FGDNet [30] | MNNet [37] | SANet (Ours) |
|---|---|---|---|---|---|---|---|---|---|
| Number of Parameters | 31.79 M | 10.45 M | 15.73 M | 88.67 M | 20.60 M | 17.20 M | 1.63 M | 0.76 M | 4.64 M |
| FLOPs | 61.56 G | 6.63 G | 31.25 G | 38.32 G | 10.24 G | 61.12 G | 6.60 G | 11.86 G | 25.26 G |
| Inference Time | 71.55 ms | 13.97 ms | 47.22 ms | 13.70 ms | 29.10 ms | 50.81 ms | 7.37 ms | 21.82 ms | 23.87 ms |

Table 5. Number of parameters, FLOPs and inference time for different denoising networks measured with input images of size $128 \times 128$.

age pairs are much larger. In this case, our SANet still outperforms all the comparative methods. Fig. 5 shows that both the single-image denoisers and the guided denoising methods FGDNet and MNNet achieve similar results where the image details are over-smoothed. In comparison, SANet can restore fine structures according to the guidance image, even in the case of large disparities. Actually, our structure aggregation strategy can also benefit other guided denoisers when tackling unaligned situations. Due to the page limit, we present the results in the supplementary materials. Fig. 6 displays the denoising results of our captured RGB-NIR stereo paired images obtained by models trained on the Flickr1024 Dataset. We can observe that SANet can be easily generalized to new cross-spectral cases. Besides, as listed in Tab. 5, SANet is also computationally efficient and has a smaller model size than those single-image denoisers.

### 4.4. Applications to Other Restoration Tasks

Our structure aggregation strategy can also handle other unaligned guided restoration tasks such as super-resolution and defocus deblurring. To validate this, we conduct experiments on the synthetic data from the Flickr1024 Dataset.

For guided super-resolution, we take the left-view image as the target and generate its low-resolution (LR) observation with 4× bicubic down-sampling. For guided deblurring, the left-view blurry target image is generated with a Gaussian kernel of standard deviation 5. The right-view image is taken as the guidance image whose channels are also switched for cross-spectral simulation. We discard the noise estimation part to predict the restored images.

In Fig. 7, two examples of a guided super-resolution result and a guided deblurring result obtained by our SANet are displayed, respectively. We also compare them to a latest single-image super-resolution model SRDD [24] and a single-image deblurring model MIMO [6]. We can observe that, compared with the single-image methods, even though the edges and details of the input target images are seriously distorted, our model can still estimate salient structure maps according to the unaligned guidance image acquired from different spectral bands. In a word, our structure aggregation is robust to various degradation factors, and thus has the potential to handle a wider range of application scenarios.

## 5. Conclusions

In this work, for the first time, we propose a guided denoising framework for cross-spectral stereo image pairs. Instead of conducting conventional stereo matching to align the guidance image with the noisy target one, we introduce a structure aggregation strategy to estimate a clean structure map from the unaligned guidance image, and then regress the final denoising result using a spatially variant linear representation model. Based on this, we further design a neural network, called as SANet, to complete the entire guided denoising process. Experimental results show that, our algorithm outperforms other state-of-the-art denoisers on various stereo datasets in both accuracy and visual quality. In addition, our structure aggregation strategy also shows its potential to handle other guided restoration tasks such as super-resolution and deblurring.

**Acknowledgement**

This work was supported in part by the "Pioneer" and "Leading Goose" R & D Program of Zhejiang under grant 2023C03136 and in part by the Ten Thousand Talents Program of Zhejiang Province under grant 2020R52003.

## References


[1] Cristhian A Aguilera, Francisco J Aguilera, Angel D Sappa, Cristhian Aguilera, and Ricardo Toledo. Learning cross-spectral similarity measures with deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2016. 3

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 60–65. IEEE, 2005. 2

[4] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. HINet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 182–192, 2021. 2, 6, 7, 8

[5] Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Shuaicheng Liu. NBNet: Noise basis learning for image denoising with subspace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4896–4906, 2021. 1, 6, 7, 8

[6] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4641–4650, 2021. 8

[7] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007. 2

[8] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3333–3348, 2020. 1, 3

[9] Weisheng Dong, Lei Zhang, and Guangming Shi. Centralized sparse representation for image restoration. In *Proceedings of the International Conference on Computer Vision*, pages 1259–1266. IEEE, 2011. 2

[10] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2862–2869, 2014. 2

[11] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1712–1722, 2019. 2

[12] Honey Gupta and Kaushik Mitra. Toward unaligned guided thermal super-resolution. *IEEE Transactions on Image Processing*, 31:433–445, 2021. 3

[13] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, 2012. 1, 2

[14] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[15] Shuangping Jin, Bingbing Yu, Minhao Jing, Yi Zhou, Jiajun Liang, and Renhe Ji. DarkVisionNet: Low-light imaging via RGB-NIR fusion with deep inconsistency prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 1, pages 1104–1112, 2022. 1

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 694–711. Springer, 2016. 5

[17] Muhammad Umar Karim Khan, Asim Khan, Jinyeon Lim, Said Hamidov, Won-Seok Choi, Woojin Yun, Yeongmin Lee, Young-Gyu Kim, Hyun-Sang Park, and Chong-Min Kyung. Offset aperture: A passive single-lens camera for depth sensing. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(5):1380–1393, 2018. 1

[18] Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N Do, and Kwanghoon Sohn. DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2103–2112, 2015. 3, 5, 7

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] Mingyang Liang, Xiaoyang Guo, Hongsheng Li, Xiaogang Wang, and You Song. Unsupervised cross-spectral stereo matching by learning to synthesize. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8706–8713, 2019. 3

[21] Hangfan Liu, Ruiqin Xiong, Dong Liu, Siwei Ma, Feng Wu, and Wen Gao. Image denoising via low rank regularization exploiting intra and inter patch correlation. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(12):3321–3332, 2017. 2

[22] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[23] Feifan Lv, Yinqiang Zheng, Yicheng Li, and Feng Lu. An integrated enhancement solution for 24-hour colorful imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11725–11732, 2020. 1

[24] Shunta Maeda. Image super-resolution with deep dictionary. In *Proceedings of the European Conference on Computer Vision*, 2022. 8

[25] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. 5

[26] Chong Mou, Qian Wang, and Jian Zhang. Deep generalized unfolding networks for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17399–17410, 2022. 1, 6, 7, 8

[27] Jinshan Pan, Jiangxin Dong, Jimmy S Ren, Liang Lin, Jinhui Tang, and Ming-Hsuan Yang. Spatially variant linear representation models for joint filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1702–1711, 2019. 1, 2, 3

[28] Di Qiu, Jiahao Pang, Wenxiu Sun, and Chengxi Yang. Deep end-to-end alignment and refinement for time-of-flight rgb-d module. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9994–10003, 2019. 3

[29] Xiaoyong Shen, Li Xu, Qi Zhang, and Jiaya Jia. Multi-modal and multi-spectral registration for natural images. In *Proceedings of the European Conference on Computer Vision*, pages 309–324. Springer, 2014. 3

[30] Zehua Sheng, Xiongwei Liu, Si-Yuan Cao, Hui-Liang Shen, and Huaqi Zhang. Frequency-domain deep guided image denoising. *IEEE Transactions on Multimedia*, 2022. 1, 2, 4, 5, 6, 7, 8

[31] Zenglin Shi, Yunlu Chen, Efstratios Gavves, Pascal Mettes, and Cees GM Snoek. Unsharp mask guided filtering. *IEEE Transactions on Image Processing*, 30:7472–7485, 2021. 2

[32] Takashi Shibata, Masayuki Tanaka, and Masatoshi Okutomi. Misalignment-robust joint filter for cross-modal image pairs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3295–3304, 2017. 3

[33] Chunwei Tian, Yong Xu, Zuoyong Li, Wangmeng Zuo, Lunke Fei, and Hong Liu. Attention-guided CNN for image denoising. *Neural Networks*, 124:117–129, 2020. 7

[34] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 5

[35] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 1, 2, 6, 7, 8

[36] Jun Xu, Lei Zhang, and David Zhang. A trilateral weighted sparse coding scheme for real-world image denoising. In *Proceedings of the European Conference on Computer Vision*, pages 20–36, 2018. 2

[37] Shuang Xu, Jiangshe Zhang, Jialin Wang, Kai Sun, Chunxia Zhang, Junmin Liu, and Junying Hu. A model-driven network for guided image denoising. *Information Fusion*, 2022. 1, 3, 6, 7, 8

[38] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 2

[39] Qiong Yan, Xiaoyong Shen, Li Xu, Shaojie Zhuo, Xiaopeng Zhang, Liang Shen, and Jiaya Jia. Cross-field joint image restoration via scale map. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1537–1544, 2013. 1, 2

[40] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Proceedings of the European Conference on Computer Vision*, pages 492–511. Springer, 2020. 2, 6, 7, 8

[41] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021. 1, 2, 6, 7, 8

[42] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 2

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

[44] Xiaopeng Zhang, Terence Sim, and Xiaoping Miao. Enhancing photographs with near infra-red images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 1

[45] Tiancheng Zhi, Bernardo R Pires, Martial Hebert, and Srinivasa G Narasimhan. Deep material-aware cross-spectral stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1916–1925, 2018. 3, 5