Structure Aggregation for Cross-Spectral Stereo Image Guided Denoising Supplementary Materials

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

In the supplementary materials, we'll provide more implementation details about frequency decomposition in the guided denoising module and more visual comparisons between our proposed SANet and competing denoising methods. In the manuscript, we have shown that the DASC-based stereo matching technique [3] achieves plausible results in the case of clean input images. However, the matching accuracy severely decreases in presence of noise. In the supplementary materials, we further warp the guidance images based on the disparity maps computed by DASC and other two learning-based unsupervised stereo matching approaches [4, 13], and demonstrate the guided denoising results using these warped guidance images. Besides, we will also show that our structure aggregation module can benefit other guided denoisers when handling unaligned situations. Finally, to better illustrate the structure aggregation process, we visualize the perceptual weights of some circularly shifted guidance images.

1. Frequency Decomposition for Guided Image Denoising

In the guided denoising module, we leverage a spatially variant linear representation model to regress the final denoising result $\hat{\mathbf{X}}$ using the structure map $\mathbf{U}$ estimated by our structure aggregation strategy. The representation model is mathematically formulated as

$$\hat{\mathbf{X}} = \mathbf{W}^S \odot \mathbf{U} + \mathbf{B}, \tag{1}$$

where $\mathbf{W}^S$ is the scale weight to adjust the structure intensity of the structure map, and $\mathbf{B}$ is the bias term to ensure that the pixel intensities are faithfully restored according to the target image. In other words, $\mathbf{W}^S$ and $\mathbf{B}$ can be regarded as focusing on representing the high-frequency and the low-frequency contents, respectively. Therefore, we learn the representation model in the frequency domain. To better perceive the low-frequency contents of the target image, the bias term $\mathbf{B}$ is estimated by a weighted fusion of the input noisy image $\mathbf{Y}$ and the estimated noise map $\hat{\mathbf{N}}$.

In this work, following [7], we also use patch-wise 2D discrete cosine transform (2D-DCT) for frequency decomposition. Denote $\mathcal{T}(\cdot)$ as the frequency decomposition function, the representation model can be re-written as

$$\mathcal{T}(\hat{\mathbf{X}}) = \mathbf{W}_T^Y \odot \mathcal{T}(\mathbf{Y}) + \mathbf{W}_T^N \odot \mathcal{T}(\hat{\mathbf{N}}) + \mathbf{W}_T^S \odot \mathcal{T}(\mathbf{U}). \tag{2}$$

Specifically, the frequency decomposition is conducted in sliding windows of size $k \times k$. For a patch $\mathbf{p}_{ij}$ centered at position (i, j) of $\mathbf{Y}$, we compute its frequency coefficients $\mathbf{q}_{ij}$ using 2D-DCT. That is,

$$\begin{aligned}\mathbf{q}_{ij}(u,v) =& \xi(u)\cdot\xi(v)\cdot\sum_{x=0}^{k-1}\sum_{y=0}^{k-1}\Bigg[\mathbf{p}_{ij}(x,y) \\ &\cdot\cos\left(\frac{(x+0.5)\pi}{k}u\right)\cdot\cos\left(\frac{(y+0.5)\pi}{k}v\right)\Bigg],\end{aligned} \tag{3}$$

where $\xi(u) = \sqrt{2/k}$ for $u = 1, \cdots, k-1$ and $\xi(0) = 1$. Then, we reshape $\mathbf{q}_{ij}$ into a vector $\tilde{\mathbf{q}}_{ij}$ of size $1 \times k^2$. Stacking the frequency coefficients of all patches together, we obtain $\mathcal{T}(\mathbf{Y}) \in \mathbb{R}^{H \times W \times k^2}$, where $\mathcal{T}(\mathbf{Y})(i, j, l) = \mathbf{q}_{ij}(l)$. $\mathcal{T}(\hat{\mathbf{N}})$ and $\mathcal{T}(\mathbf{U})$ are obtained in the same way. In addition, since 2D-DCT is computed with a set of fixed and spatially-invariant linear coefficients, it can be easily implemented using a convolution layer with k^2 fixed DCT kernels of size $k \times k$ for GPU acceleration. Similarly, to transform the frequency coefficients of the denoised image back to the spatial domain, the inverse 2D-DCT can also be implemented using a convolution layer. In [7], the authors provide the implementation codes for the 2D-DCT and the inverse 2D-DCT convolution layers.

2. Additional Experimental Results on Cross-Spectral Stereo Image Pairs

In the manuscript, we show that it's quite challenging to guarantee the pixel-level registration accuracy for current

*Corresponding author.

Methods	$\sigma = 0.2$			$\alpha = 0.02, \sigma = 0.2$		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
w/o structure aggregation	25.30	0.8345	0.2945	24.91	0.8266	0.3008
w/ warped guidance image (DASC [3])	25.38	0.8379	0.2934	24.99	0.8286	0.3054
w/ warped guidance image (Zhi *et al*. [13])	25.34	0.8360	0.2952	24.96	0.8270	0.3085
w/ warped guidance image (Liang *et al*. [4])	25.43	0.8403	0.2867	24.99	0.8285	0.3004
w/ structure aggregation (ours)	**25.67**	**0.8477**	**0.2685**	**25.30**	**0.8411**	**0.2736**

Table 1. The average PSNR (dB), SSIM and LPIPS values of denoising results obtained using the warped guidance images by DASC [3], Zhi *et al*. [13], Liang *et al*. [4] and our estimated structure maps on images from the Flickr1024 Dataset [8] under Gaussian noise ($\sigma = 0.2$) and mixed Poisson-Gaussian noise ($\alpha = 0.02$, $\sigma = 0.2$).

Methods	FGDNet [7]						MNNet [10]					
	$\sigma = 0.2$			$\alpha = 0.02, \sigma = 0.2$			$\sigma = 0.2$			$\alpha = 0.02, \sigma = 0.2$		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Original	24.97	0.8185	0.3107	24.54	0.8102	0.3189	25.12	0.8269	0.3181	24.73	0.8173	0.3259
w/ DASC [3]	25.13	0.8291	0.3171	24.74	0.8194	0.3279	25.21	0.8314	0.3062	24.82	0.8223	0.3194
w/ Zhi *et al*. [13]	25.13	0.8288	0.3174	24.71	0.8172	0.3302	25.18	0.8283	0.3094	24.78	0.8174	0.3293
w/ Liang *et al*. [4]	25.14	0.8293	0.3139	24.76	0.8207	0.3247	25.23	0.8312	0.3111	24.68	0.8216	0.3081
w/ SA (ours)	**25.23**	**0.8337**	**0.2846**	**24.83**	**0.8264**	**0.2923**	**25.66**	**0.8456**	**0.2799**	**25.30**	**0.8367**	**0.2867**

Table 2. Evaluation results on other guided denoisers with different matching algorithms and our structure aggregation (SA) module on the Flickr1024 Dataset under Gaussian noise ($\sigma = 0.2$) and mixed Poisson-Gaussian noise ($\alpha = 0.02$, $\sigma = 0.2$).

cross-spectral stereo matching algorithms, especially in the presence of noise. Tab. 1 lists the quantitative denoising results obtained using the warped guidance images computed by the DASC-based stereo matching technique [3] and two state-of-the-art unsupervised cross-modal stereo matching networks [4, 13]. We can observe that they basically achieve very similar PSNR, SSIM and LPIPS values than those obtained using the original unaligned guidance images. Therefore, conventional stereo matching does not bring much improvement to the denoising performance. Visual comparisons in Fig. 1 display that an inaccurate warped guidance image cannot solve the problem of over-smoothing weak details in the guided denoising process. In comparison, our structure aggregation strategy can produce a structure map that is structurally aligned with the input target image before the edges and details can be effectively transferred to the denoising result.

To further show the effectiveness of our proposed structure aggregation strategy, we pre-align the input image pairs using different stereo matching methods [3, 4, 13] and our structure aggregation model. Then, we perform guided denoising with FGDNet [7] and MNNet [10] that are designed for aligned situations. The quantitative results are listed in Tab. 2. We can observe that with our structure aggregation, both comparative guided denoisers achieve noticeable performance gain, better than using other alignment methods. Hence, our structure aggregation module can be regarded as a plug-and-play component that can allow previous guided denoising methods to be able to deal with more general situations.

We evaluate the denoising performance of our SANet and compare it to the state-of-the-art single-image denoisers including MIRNet [11], NBNet [2], MPRNet [12], HINet [1], Uformer [9] and DGUNet [6], as well as guided denoising methods including FGDNet [7] and MNNet [10] on the PittsStereo-RGBNIR Dataset [13], the Flickr1024 Dataset [8], and the KITTI Stereo 2015 Dataset [5]. More visual comparisons are displayed in Fig. 2-Fig. 8. As described in the manuscript, the target and the guidance images from the PittsStereo-RGB Dataset are captured in the visible and the near-infrared (NIR) bands, respectively. Considering that they basically have very small disparities, we further evaluate our proposed algorithm in more challenging situations on the Flickr1024 and the KITTI Stereo 2015 Datasets where the input paired images have much larger disparities. To simulate the cross-spectral cases, the target and the guidance images are extracted from different channels of the RGB images. In addition, we also construct an RGB-NIR stereo dual-camera system and evaluate the algorithms on realistic image pairs. The corresponding visual results are displayed in Fig. 4. We can observe that, both single-image denoisers and guided denoising models FGDNet and MNNet inevitably over-smooth detailed contents during noise removal. In comparison, our proposed SANet can effectively restore more salient structures and richer details according to the unaligned guidance images.

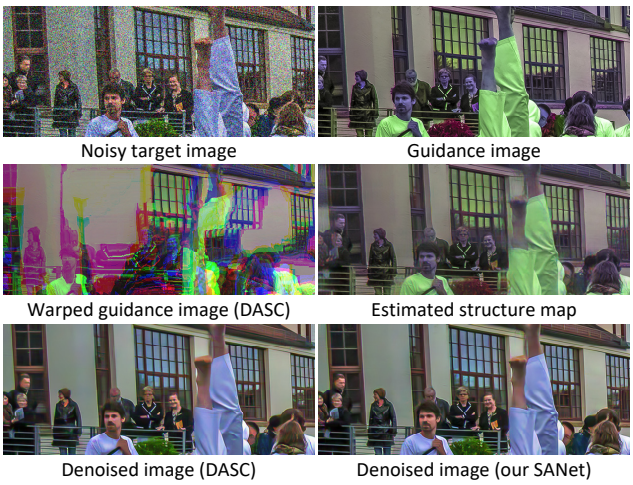

Figure 1. Guided denoising results obtained using the warped guidance image by DASC [3] and our estimated structure map under mixed Poisson-Gaussian noise ($\alpha = 0.02$, $\sigma = 0.2$).

3. Visualization of Structure Aggregation

To better demonstrate the structure aggregation process, we visualize the perceptual weights for a potion of the circularly shifted guidance images. Denote $\mathbf{G}_d \in \mathbb{R}^{H\times W}$, where $d = 0, 1, \cdots, D$ is the shifted distance. For the d-th shifted guidance image, its corresponding perceptual weight is denoted as $\mathbf{W}_d^P \in \mathbb{R}^{H\times W}$. Therefore, $\mathbf{G}_d \odot \mathbf{W}_d^P$ can be regarded as the perceived consistent structures, where $\odot$ is the element-wise product operator.

Fig. 9 displays the visualization results on our captured RGB-NIR stereo image pairs. The estimated structure map is obtained by our SANet trained using the synthetic cross-spectral image pairs from the Flickr1024 Dataset. We can observe that, even if the test data are captured with different cross-spectral settings, our proposed SANet still performs well, demonstrating its good generalizability. Taking the blue channel of the noisy target image as an example, the structure map is estimated based on an NIR guidance image captured with a stereo dual-camera system. Without an explicit matching process, our structure aggregation strategy can effectively perceive the structural correlation between the noisy target image and the shifted guidance image, and extracts consistent contents to synthesize the structure map.

References

[1] 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

[2] 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. 2

[3] 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. 1, 2, 3

[4] 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. 1, 2

[5] 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. 2

[6] 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. 2

[7] 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

[8] 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. 2

[9] 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. 2

[10] 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. 2

[11] 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

[12] 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. 2

[13] 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. 1, 2

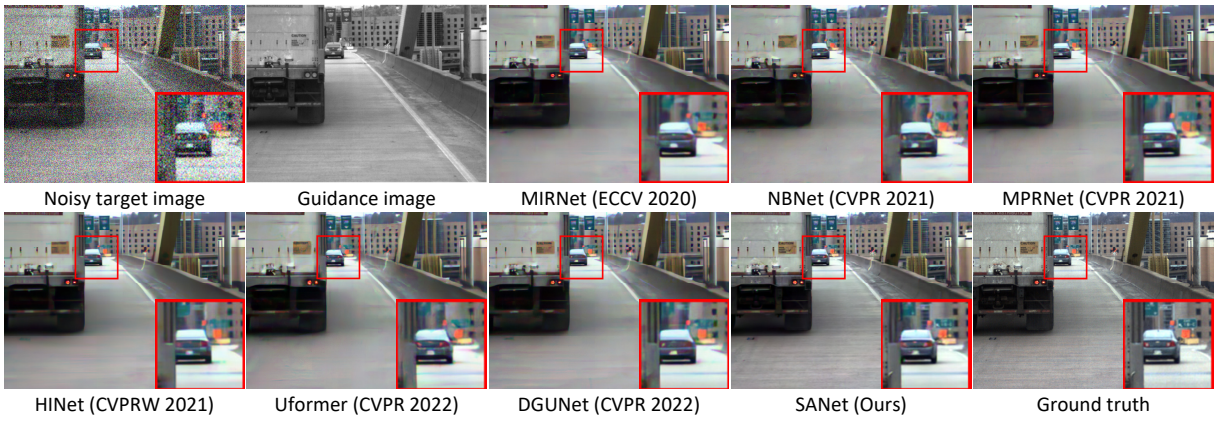

Figure 2. Denoising results on the PittsStereo-RGBNIR Dataset under Gaussian noise ($\sigma = 0.2$) obtained by the comparative denoising methods and our SANet.

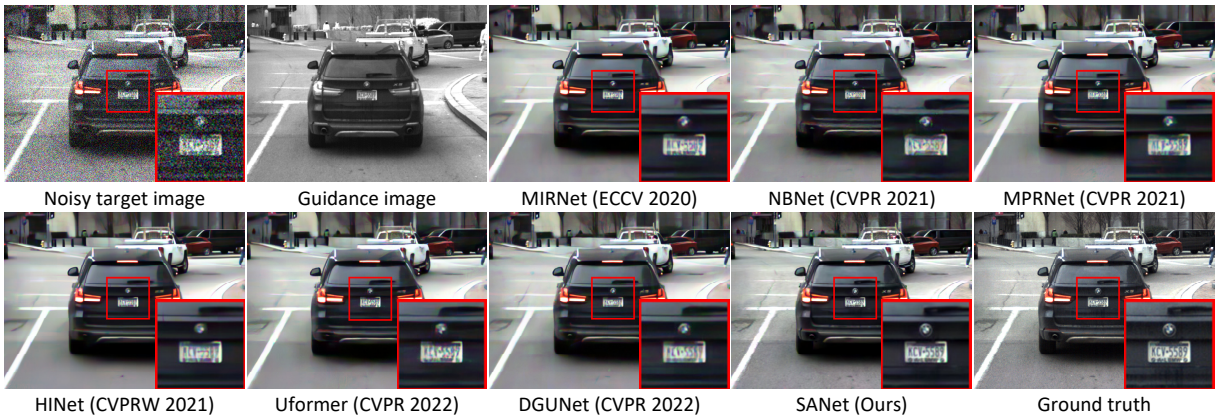

Figure 3. Denoising results on the PittsStereo-RGBNIR Dataset under mixed Poisson-Gaussian noise ($\alpha = 0.02$, $\sigma = 0.2$) obtained by the comparative denoising methods and our SANet.

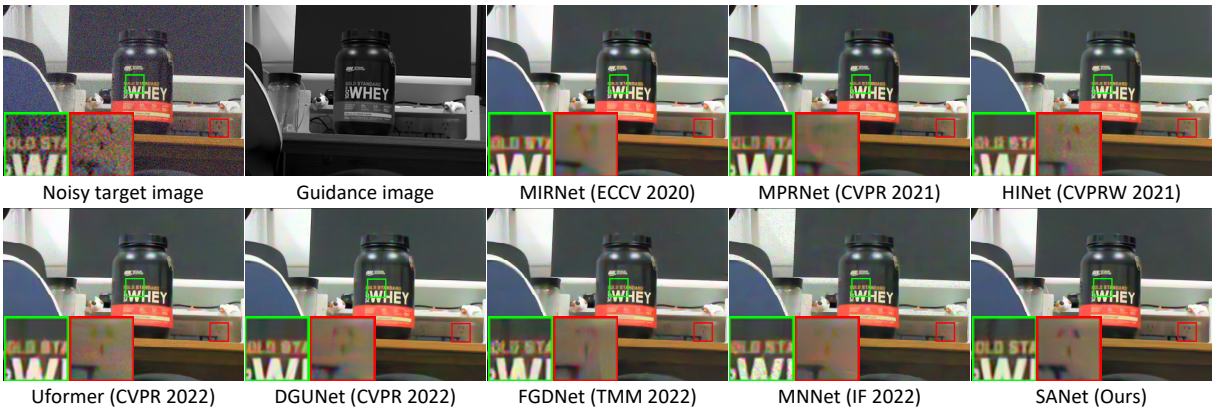

Figure 4. Denoising results of our captured realistic RGB-NIR image pair obtained by the comparative denoising methods and out SANet. All images are processed with tone mapping for better illustration.

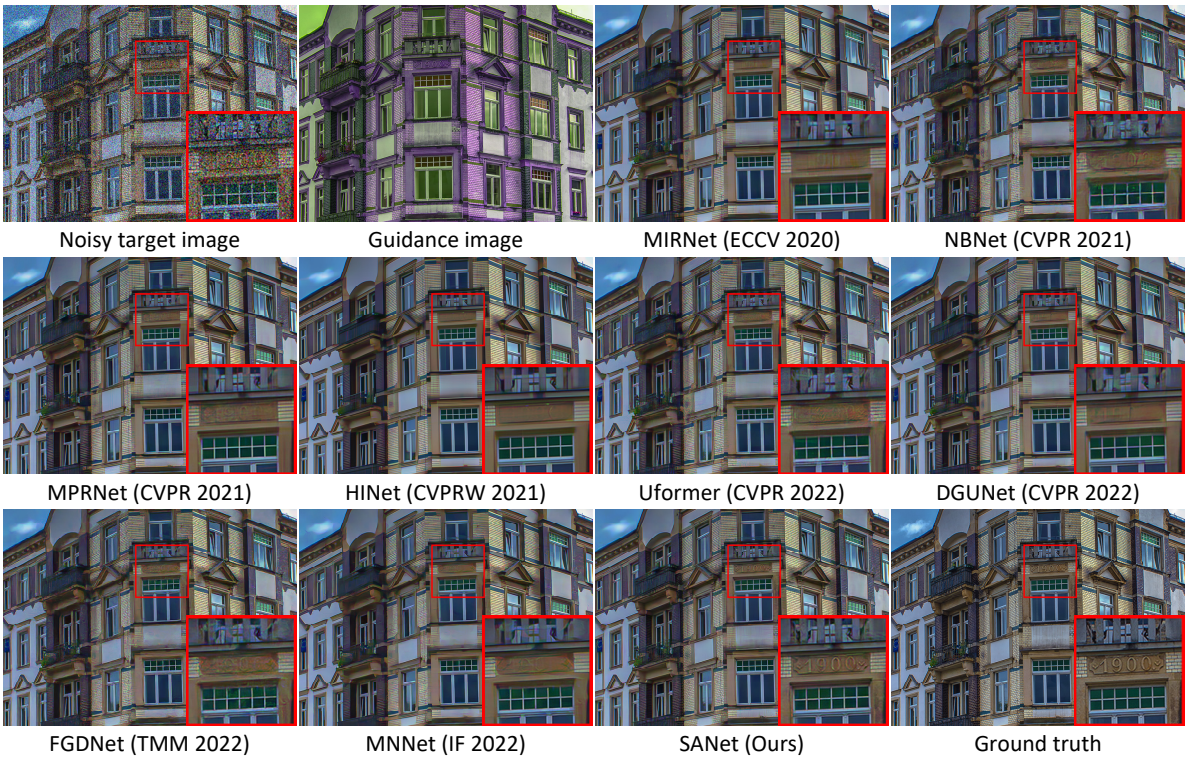

Figure 5. Denoising results on the Flickr1024 Dataset under mixed Poisson-Gaussian noise ($\alpha = 0.02$, $\sigma = 0.2$) obtained by the comparative denoising methods and our SANet.

Figure 6. Denoising results on the Flickr1024 Dataset under Gaussian noise ($\sigma = 0.2$) obtained by the comparative denoising methods and our SANet.

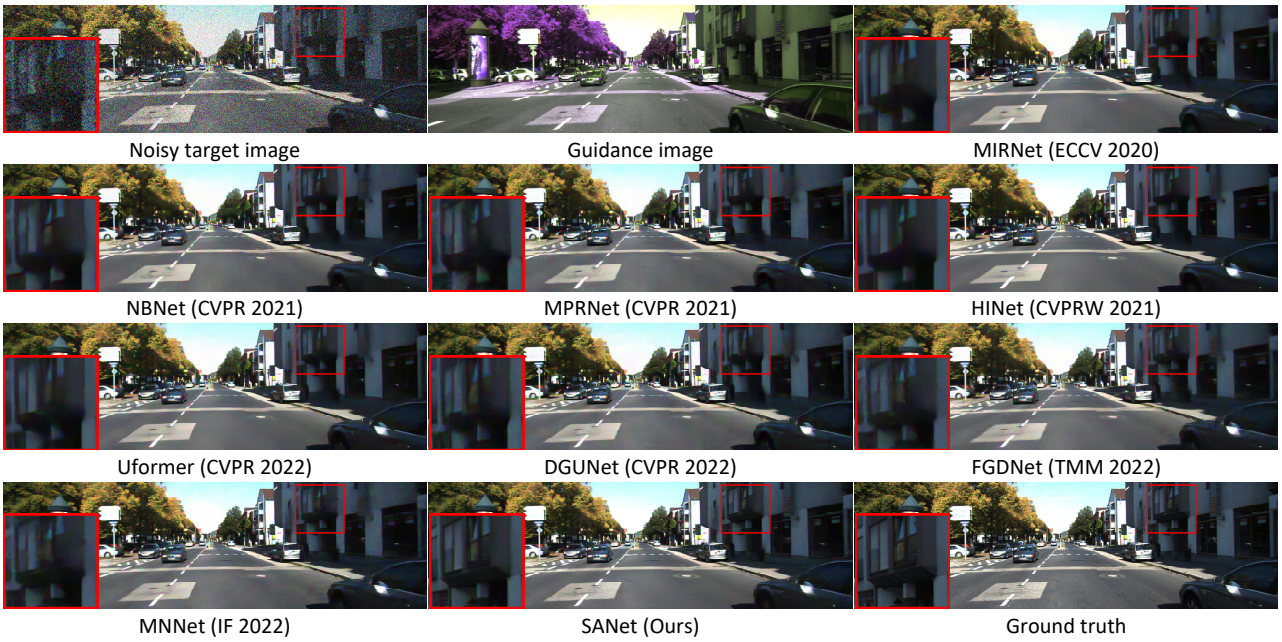

Figure 7. Denoising results on the KITTI Stereo 2015 Dataset under mixed Poisson-Gaussian noise ($\alpha = 0.02$, $\sigma = 0.2$) obtained by the comparative denoising methods and our SANet.

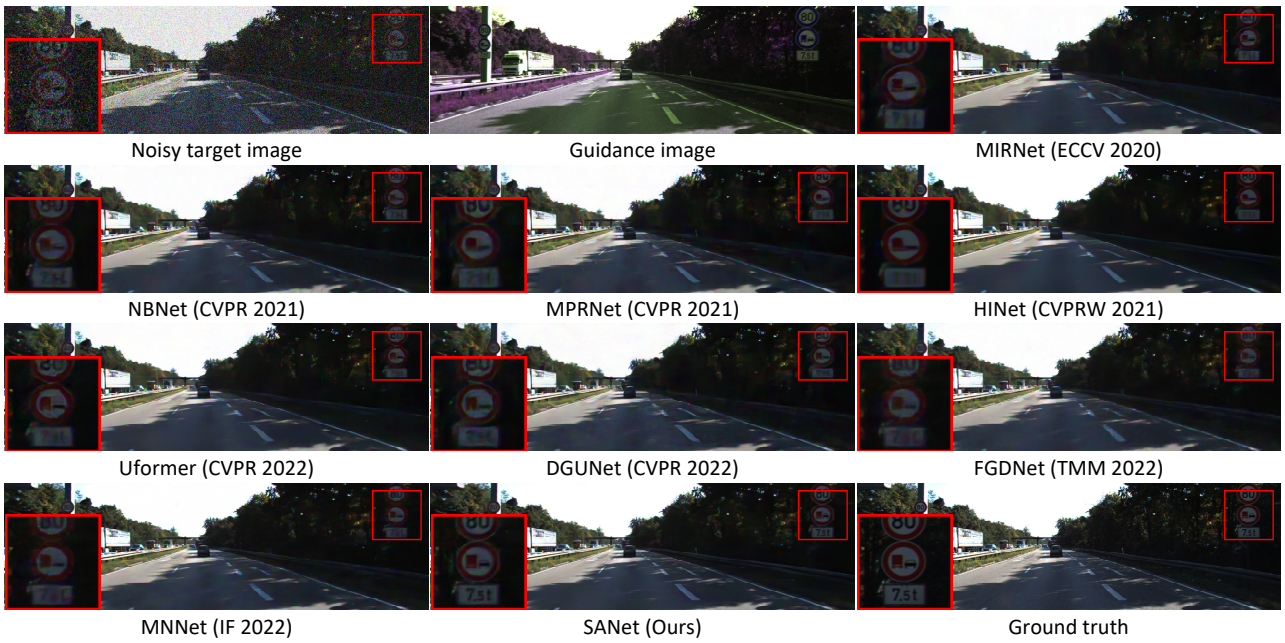

Figure 8. Denoising results on the KITTI Stereo 2015 Dataset under mixed Poisson-Gaussian noise ($\alpha = 0.02$, $\sigma = 0.2$) obtained by the comparative denoising methods and our SANet.

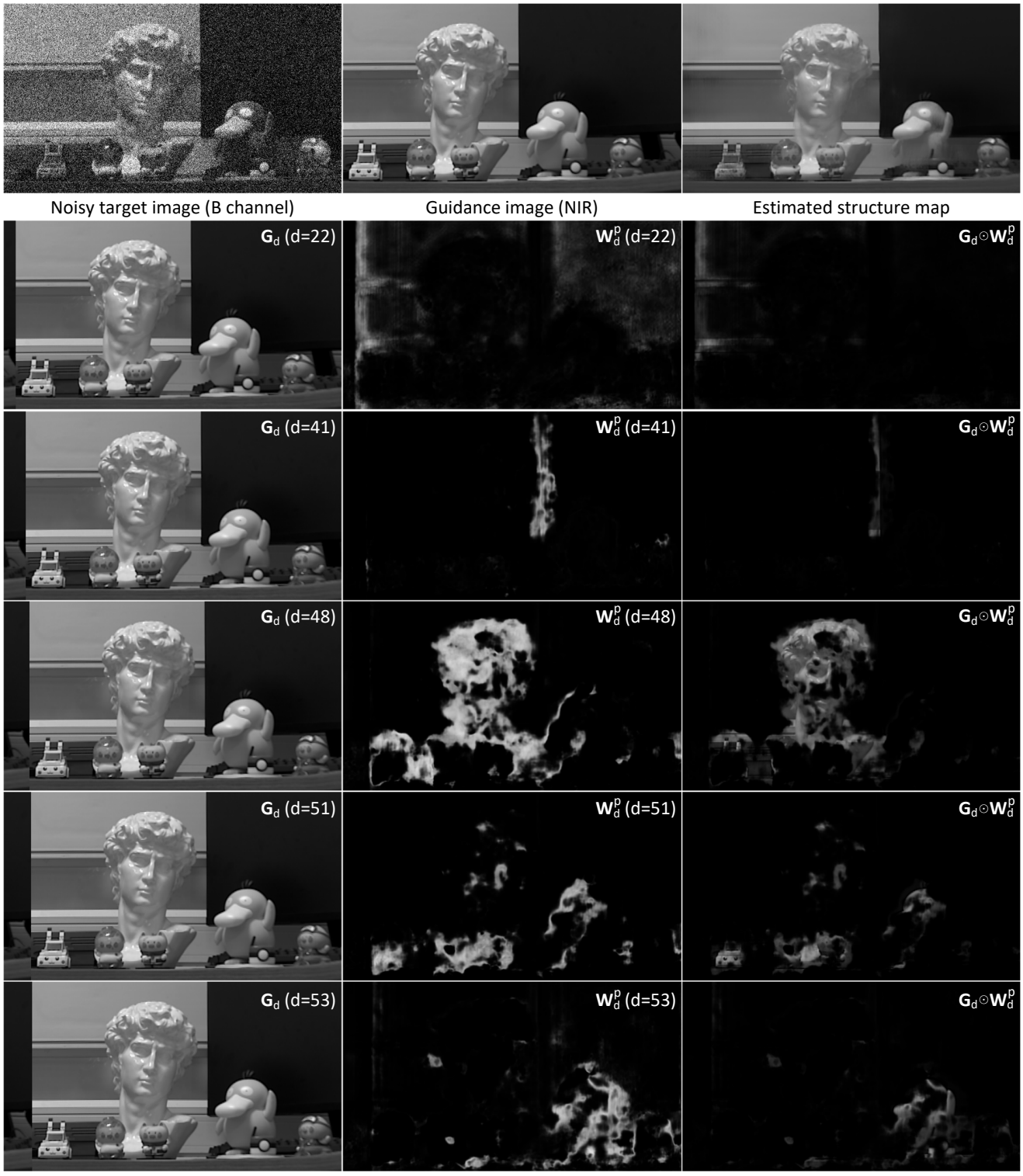

Figure 9. Visualization of our structure aggregation process, including the shifted guidance images $\{\mathbf{G}_d\}_{d=0,1,\cdots,D}$ and the corresponding perceptual weights $\{\mathbf{W}_d^P\}_{d=0,1,\cdots,D}$.