Deep Denoising of Flash and No-Flash Pairs for Photography in Low-Light Environments Supplementary Material

Zhihao Xia[1], Michaël Gharbi[2], Federico Perazzi[3], Kalyan Sunkavalli[2], Ayan Chakrabarti[1]
[1]Washington University in St. Louis [2]Adobe Research [3]Facebook
{zhihao.xia,ayan}@wustl.edu, {mgharbi,sunkaval}@adobe.com, fperazzi@fb.com

A. Architecture Details

We describe our network architecture in detail in Table 3. Our architecture follows the encoder with dual decoder architecture of [35], but changes the output of the global decoder to output a basis with the two sets of kernels $\{(A_j, B_j)\}$, each a $K \times K \times 3$ kernel, and the per-pixel decoder outputs a scale map in addition to the kernel coefficients.

The output of the global decoder gives us $J = 90$ pairs of kernels $\{(A_j, B_j)\}$, and that of the per-pixel decoder both the coefficients $C[n] \in \mathbb{R}^J$ and a scale map $G[n] \in \mathbb{R}^3$. As noted in the paper, we set $J = 90$ in our experiments.

Baselines. Our baselines use similar architectures to our main method to enable a fair comparison. For the no-flash single image input, we use the above architecture to take only the no-flash image as input (6 channels: 3 for the image itself and 3 for noise deviation maps), and do not output a scale map (i.e., our per-pixel decoder only outputs the J channel coefficient map). For the BPN [35] entries in our table for the 2x burst of no-flash images as well as for flash and no-flash denoising, we use the original architecture from their paper. Specifically, our per-pixel decoder again does not output a scale map, and the two sets of kernels are used to filter the two input images which are then added—unlike our approach which combines the two kernels to create a larger upsampled kernel that is only applied to the ambient, followed by multiplication with the scale map. For the direct prediction and KPN [27], we use this architecture without the global decoder. For direct prediction, the per-pixel decoder just outputs a 3-channel map that is treated as a residual and added to the noisy no-flash input to yield the final denoised output. For KPN, the per-pixel decoder outputs a 150-channel output: these are interpreted as two 5×5 kernels per color channel, to be applied to the flash and no-flash pairs.

B. Additional Results

Qualitative Results. We show comparison results on more images in Figure 7.

No-Flash vs. Flash as reference. As noted in the paper, using the no-flash image as the geometric reference leads to better performance at all but the darkest light level. This is true not just for our approach, but also the other baselines we considered for denoising with flash and no-flash image pairs as input. We report the performance of these methods when using flash as reference in Table 4, and find that like in the case of no-flash reference, our approach yields superior reconstructions. Although slightly worse on average, we find that using flash as reference can sometimes lead to superior reconstruction of high-frequency details for some images compared to the no-flash reference, and even when the results are quantitatively worse, this is because of misalignment of low-frequency shading that is often not perceptible. We illustrate this with qualitative comparisons in Fig. 8.

Other Noise Levels. In addition to the original values of read and shot noise variances used in the main results table, Table 5 reports the performance of our method, and other baselines for denoising flash and no-flash pairs, for three additional sets of noise parameters at one of the light levels.

Burst Denoising. We compared to using a burst of two no-flash images in the paper, as a means of evaluating the relative benefit of a the second image being taken with vs. without a flash. In both cases the second image provides additional information—a second no-flash image has high noise (though a different realization of noise than the first image), while a

Name	Input	Layer	Output Size
Input	-	-	H x W x 12
Encoder			
Enc-0	Input	3x3 Conv	H x W x 64
Enc-1-A	Enc-0	3x3 Conv	H x W x 64
Enc-1-B	Enc-1-A	3x3 Conv	H x W x 64
Enc-1-C	Enc-1-B	2x2 Stride 2 Max Pool	H/2 x W/2 x 64
Enc-2-A	Enc-1-C	3x3 Conv	H/2 x W/2 x 128
Enc-2-B	Enc-2-A	3x3 Conv	H/2 x W/2 x 128
Enc-2-C	Enc-2-B	2x2 Stride 2 Max Pool	H/4 x W/4 x 128
Enc-3-A	Enc-2-C	3x3 Conv	H/4 x W/4 x 256
Enc-3-B	Enc-3-A	3x3 Conv	H/4 x W/4 x 256
Enc-3-C	Enc-3-B	2x2 Stride 2 Max Pool	H/8 x W/8 x 256
Enc-4-A	Enc-3-C	3x3 Conv	H/8 x W/8 x 512
Enc-4-B	Enc-4-A	3x3 Conv	H/8 x W/8 x 512
Enc-4-C	Enc-4-B	2x2 Stride 2 Max Pool	H/16 x W/16 x 512
Enc-5-A	Enc-4-C	3x3 Conv	H/16 x W/16 x 1024
Enc-5-B	Enc-5-A	3x3 Conv	H/16 x W/16 x 1024
Enc-5-C	Enc-5-B	2x2 Stride 2 Max Pool	H/32 x W/32 x 1024
Enc-F	Enc-5-C	3x3 Conv	H/32 x W/32 x 1024
Enc-Out	Enc-F	3x3 Conv	H/32 x W/32 x 1024
Global Decoder			
GDec-5-A	Bilinear-Up(GP(Enc-Out))	3x3 Conv	2x2 x 512
GDec-5-B	GDec-5-A, GP-R[2x2](Enc-5-B)	3x3 Conv	2 x 2 x 512
GDec-5-C	GDec-5-B	3x3 Conv	2 x 2 x 512
GDec-4-A	Bilinear-Up(GDec-5-C)	3x3 Conv	4 x 4 x 256
GDec-4-B	GDec-4-A, GP-R[4x4](Enc-4-B)	3x3 Conv	4 x 4 x 256
GDec-4-C	GDec-4-B	3x3 Conv	4 x 4 x 256
GDec-3-A	Bilinear-Up(GDec-4C)	3x3 Conv	8 x 8 x 256
GDec-3-B	GDec-3-A, GP-R[8,8](Enc-3-B)	3x3 Conv	8 x 8 x 256
GDec-3-C	GDec-3-B	3x3 Conv	8 x 8 x 256
GDec-2-A	Bilinear-Up(GDec-3-C)	3x3 Conv	16 x 16 x 128
GDec-2-B	GDec-2-A, GP-R[16,16](Enc-2-B)	3x3 Conv	16 x 16 x 128
GDec-2-C	GDec-2-B	3x3 Conv	16 x 16 x 128
GDec-F-A	GDec-2-C	2x2 Conv (Valid)	15 x 15 x 128
GDec-F-B	GDec-F-A	3x3 Conv	15 x 15 x 128
Output: Basis	GDec-F-B	3x3 Conv	15 x 15 x (3*2*J)

Table 3: **Our network architecture.** Bi-linear upsampling refers to upsampling the feature map by a factor of 2. GP refers to global average pooling, and GP-R[H', W'] to global average pooling followed by replicating spatially to size H' x W'. Multiple inputs are concatenated along the channel dimension before being passed to the convolution layer. All convolution layers use same padding unless otherwise specified.

flash image has a much higher signal-to-noise ratio but entirely different shading. Our results showed that in this context, a second image is beneficial.

But more generally, burst photography (which typically involves a larger number of images) has its own relative strengths and weaknesses when compared to using flash and no-flash pairs, as a means of imaging in low-light. Burst denoising with longer bursts may well be a preferable option in the presence of moderate motion, or when using a flash is not an option (for example, when most objects in the scene are far away and can not be illuminated with a flash). Conversely, the use of a flash and no-flash pair is preferable in much lower light, in scenes where most of the scene *can* be well-illuminated with a flash, when camera or scene motion may cause significant misalignment across a large sequence of images, or when memory or

Name	Input	Layer	Output Size
Per-pixel Decoder			
PDec-5-A	Bilinear-Up(Enc-Out)	3x3 Conv	H/16 x W/16 x 512
PDec-5-B	PDec-5-A, Enc-5-B	3x3 Conv	H/16 x W/16 x 512
PDec-5-C	PDec-5-B	3x3 Conv	H/16 x W/16 x 512
PDec-4-A	Bilinear-Up(PDec-5-C)	3x3 Conv	H/8 x W/8 x 256
PDec-4-B	PDec-4-A, Enc-4-B	3x3 Conv	H/8 x W/8 x 256
PDec-4-C	PDec-4-B	3x3 Conv	H/8 x W/8 x 256
PDec-3-A	Bilinear-Up(PDec-4-C)	3x3 Conv	H/4 x W/4 x 128
PDec-3-B	PDec-3-A, Enc-3-B	3x3 Conv	H/4 x W/4 x 128
PDec-3-C	PDec-3-B	3x3 Conv	H/4 x W/4 x 128
PDec-2-A	Bilinear-Up(PDec-3-C)	3x3 Conv	H/2 x W/2 x 64
PDec-2-B	PDec-2-A, Enc-2-B	3x3 Conv	H/2 x W/2 x 64
PDec-2-C	PDec-2-B	3x3 Conv	H/2 x W/2 x 64
PDec-1-A	Bilinear-Up(PDec-2-C)	3x3 Conv	H x W x 64
PDec-1-B	PDec-1-A, Enc-1-B	3x3 Conv	H x W x 64
PDec-1-C	PDec-1-B	3x3 Conv	H x W x 64
PDec-F-0	PDec-1-C	3x3 Conv	H x W x 64
PDec-F-1	PDec-F-0	3x3 Conv	H x W x 64
Output: Coeffs + Scale map	PDec-F-1	3x3 Conv	H x W x (J + 3)

Table 3: (continued) **Our network architecture.**

Method	**100x Dimmed**		**50x Dimmed**		**25x Dimmed**		**12.5x Dimmed**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Direct Prediction	24.15 dB	0.779	26.00 dB	0.810	27.31 dB	0.836	28.17 dB	0.856
KPN [27]	25.51 dB	0.820	27.43 dB	0.852	28.78 dB	0.874	29.74 dB	0.889
BPN [35]	26.23 dB	0.831	27.83 dB	0.857	29.08 dB	0.877	30.00 dB	0.891
Ours	**26.83 dB**	**0.843**	**28.39 dB**	**0.866**	**29.55 dB**	**0.883**	**30.45 dB**	**0.897**

Table 4: **Quantitative Results using flash image as geometric reference for all methods.**

Noise Parameters	$\log([\sigma_r, \sigma_s])$ = [-2.8, -4.0]		$\log([\sigma_r, \sigma_s])$ = [-2.4, -3.2]		$\log([\sigma_r, \sigma_s])$ = [-2.2, -2.8]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Direct Prediction	28.44 dB	0.845	25.54 dB	0.787	23.88 dB	0.754
KPN [27]	29.01 dB	0.870	26.82 dB	0.832	25.59 dB	0.808
BPN [35]	29.18 dB	0.870	26.86 dB	0.829	25.61 dB	0.805
Ours	**29.65 dB**	**0.876**	**27.47 dB**	**0.842**	**26.26 dB**	**0.821**

Table 5: **Noise levels.** Performance of different approaches to denoising flash and no-flash pairs, at 50x dimmed light levels, with three additional noise levels (increasing from left to right). Note that the noise level used in the main results in the paper was between the first and second level above.

computational constraints prevent capturing a larger sequence of images.

While the question of what acquisition strategy to use will depend on the environment and platform and is beyond the scope of this paper, we present a comparison in Table 6 to provide further intuition to the reader. We compare burst denoising, using BPN [35], with larger bursts of 5 and 7 images on our dataset—we use the same noise and dimming models to generate a larger burst of no-flash images. These sequences are mis-aligned using our randomly sampled homographies, with the homographies applied sequentially—thus the first and last image of a sequence will have a larger misalignment on average than two subsequent pairs, and so we use the image in the middle of the sequence as reference. We compare these results to

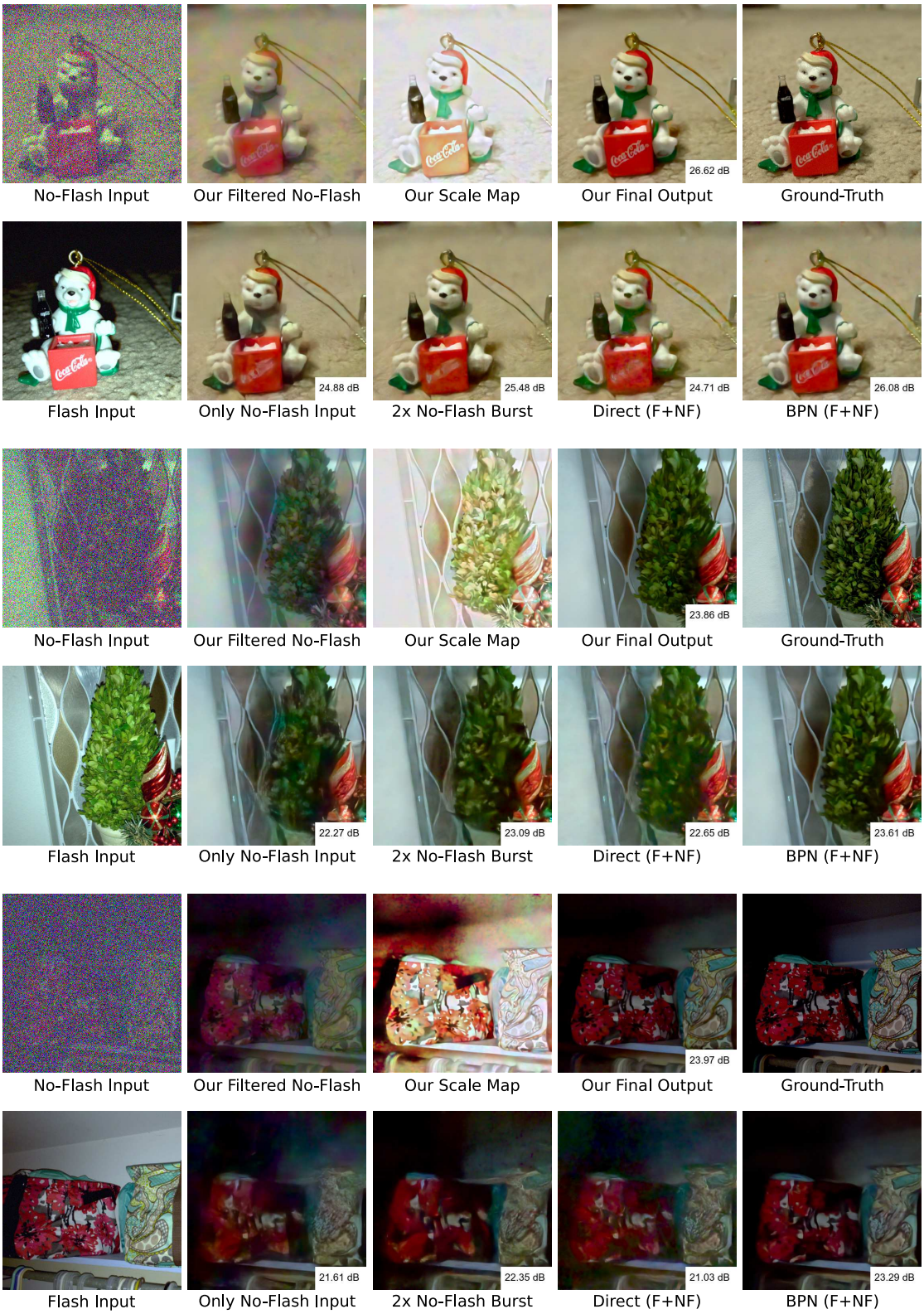

Figure 7: **More qualitative comparisons.**

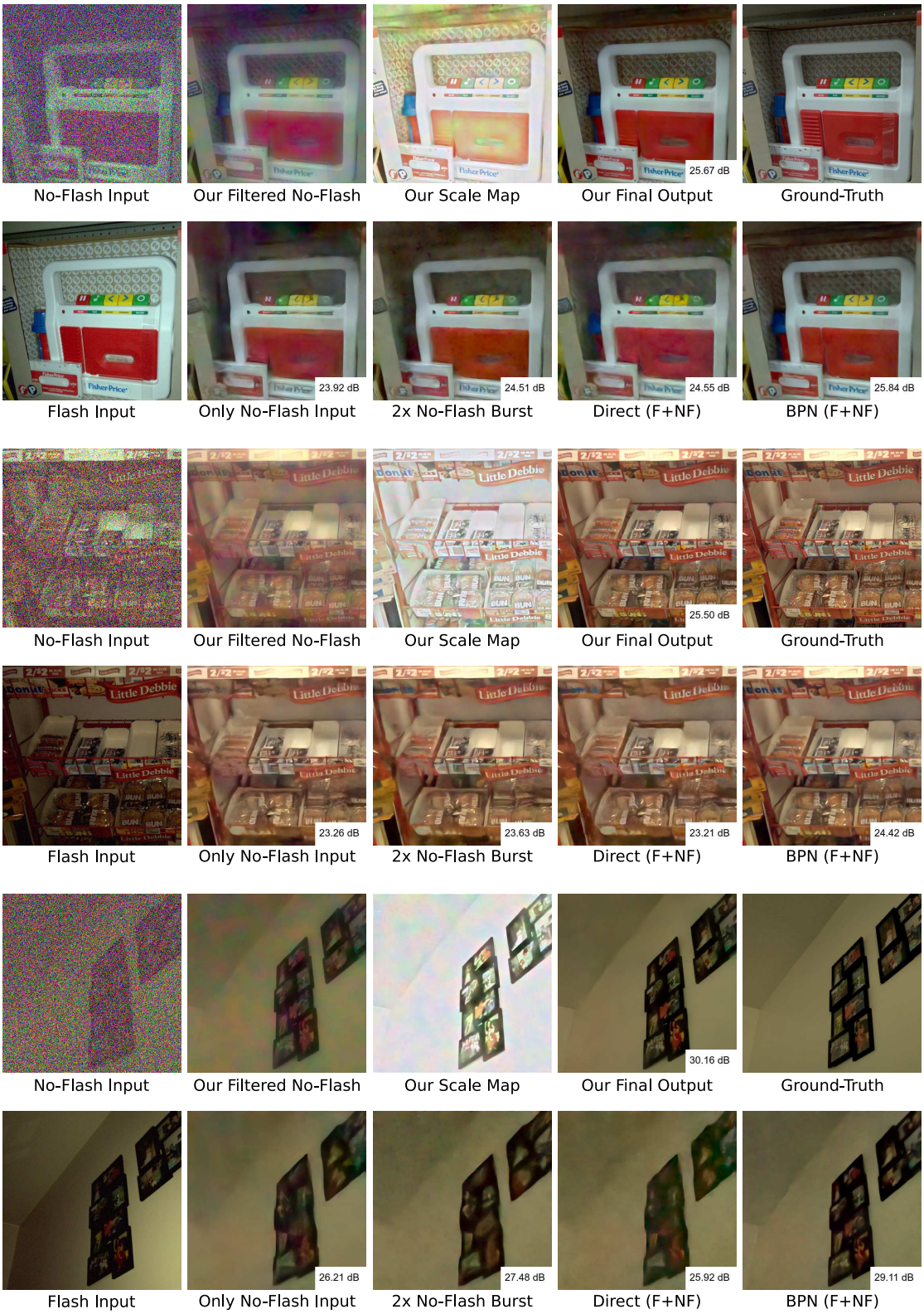

Figure 7: (continued) **More qualitative comparisons.**

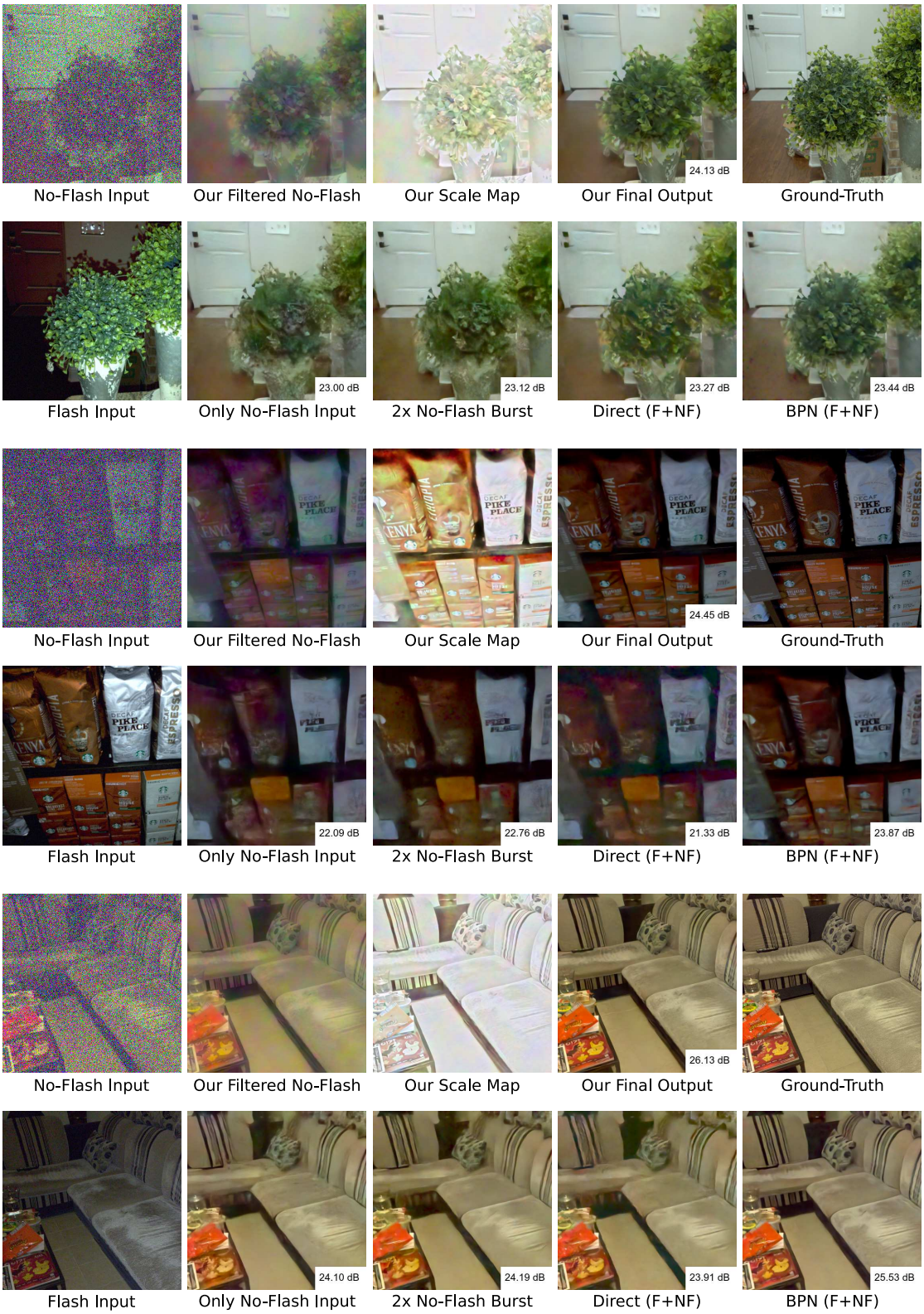

Figure 7: (continued) **More qualitative comparisons.**

Method	100x Dimmed		50x Dimmed		25x Dimmed		12.5x Dimmed	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
2x No-flash (BPN [35])	25.58 dB	0.796	27.75 dB	0.839	29.65 dB	0.874	31.21 dB	0.899
Flash and No-flash (Ours)	26.75 dB	0.829	28.56 dB	0.860	30.14 dB	0.884	31.52 dB	0.903
5x No-flash (BPN [35])	25.90 dB	0.788	27.84 dB	0.832	29.54 dB	0.867	31.03 dB	0.894
Flash and 4x No-flash (Ours)	26.82 dB	0.822	28.60 dB	0.856	30.21 dB	0.883	31.60 dB	0.904
7x No-flash (BPN [35])	26.00 dB	0.792	27.89 dB	0.834	29.57 dB	0.868	31.05 dB	0.894
Flash and 6x No-flash (Ours)	26.80 dB	0.820	28.60 dB	0.855	30.20 dB	0.882	31.59 dB	0.903

Table 6: **Performance with larger bursts.** We compare the original results of 2x no-flash burst with BPN [35] and flash and no-flash denoising with our method, to denoising larger bursts of no-flash images of length 5 and 7 with BPN, as well as using a modified version of our method with bursts of the same length where one of the images is captured with a flash.

using our method when denoising bursts of the same size, where one (the last) image is taken with a flash and the rest without (again using the middle no-flash image as reference). Here, our network predicts kernels to be used to filter and sum all the no-flash images, which is then multiplied with our scale map. Because memory constraints, we do not use kernel upsampling in these experiments, and predict only a basis of 15×15 kernels (one for each channel of each no-flash image).

Our results show that in the light and motion settings we consider, larger no-flash bursts only have a modest improvement over a pair of two no-flash images at lower light levels, although they perform slightly worse comparatively at higher light levels (this is likely because the networks are trained over a range of light levels, and tend to oversmooth to handle the lowest end of that range). Our method, when using a burst of the same size with one as a flash image, performs better than pure no-flash bursts, but also with only modest improvements over a flash and no-flash pair (note that in this case, the misalignment between the flash image and reference frame is greater than for a flash and no-flash pair that are taken in sequence). These results suggest that when using burst photography in settings where it is advantageous, it may be worth capturing one image of that burst with a flash.

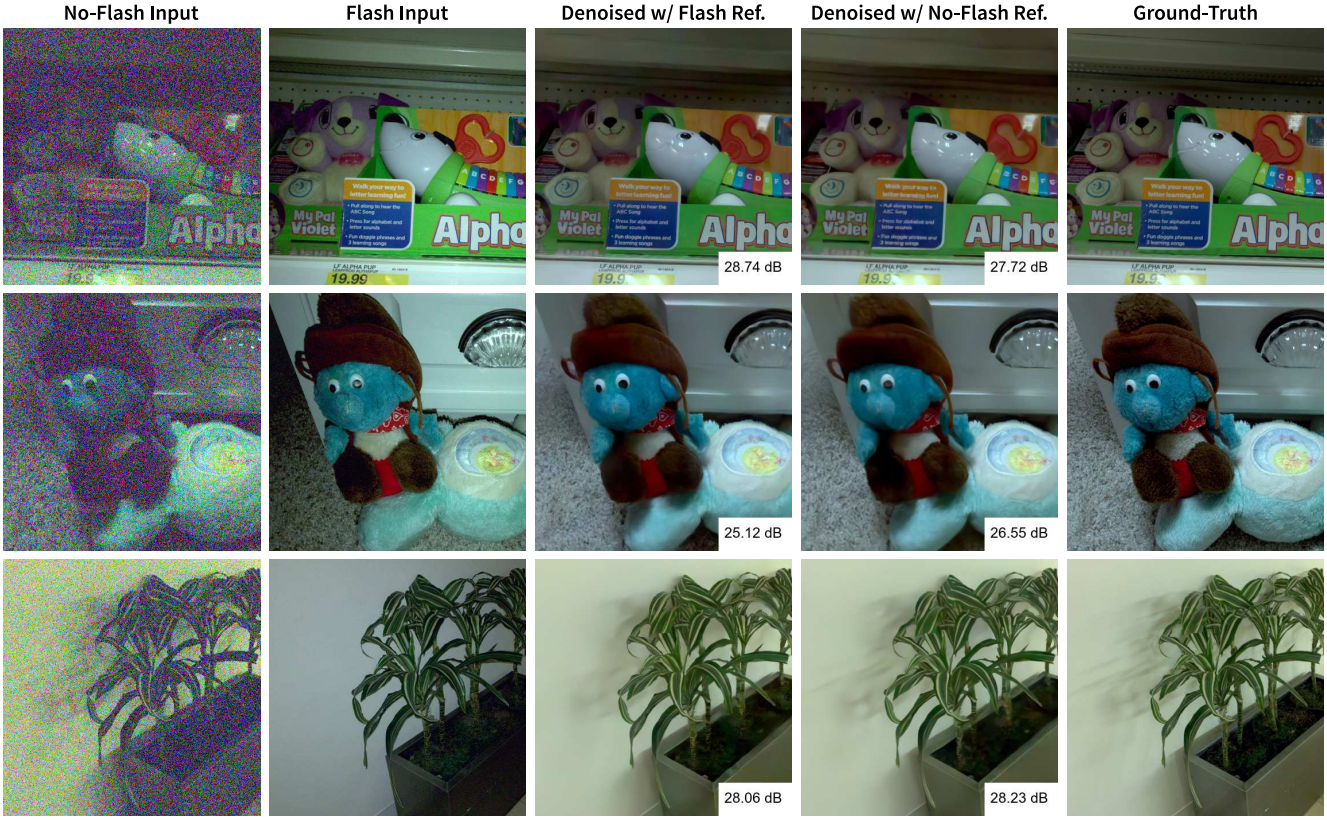

Figure 8: **Flash vs No-Flash as reference.** We show qualitative comparisons of results (at 50x dimmed light level) for our method using the flash input vs. the no-flash input as geometric reference. While using the no-flash input as reference does better on average, in some examples, using the flash as reference can lead to better reconstruction of high frequency detail (first row). In other cases, even though the flash-reference results are quantitatively, the difference is due to errors in reconstructing shading which are often less perceptually obvious. This is the case in the bottom two rows, although in the last row, we can see that the flash reference output has a blurrier reconstruction of the shadows on the wall.