Gain-first or Exposure-first: Benchmark for Better Low-light Video Photography and Enhancement

Supplementary Material

In this supplementary document, we provide detailed settings for hyper-parameters in our experiments along with computational cost of each model, a theoretical explanation for our experiment results, additional evaluations and visualizations, as well as a summary of all other resources that are available on our Project page.

A. Hyper-parameter Settings and Computational Costs

For RVRT [9], training inputs have the same spatial dimension, 256 pixels, as original paper, while temporal length is scaled down from 16 frames to 8, due to memory limits. Batch size is set to 4, with iteration number and initial learning rate remaining unchanged from the paper.

For BasicVSR++ [4], although slightly different hyper-parameter settings are used for denoising and deblurring datasets in the original paper [5], considering possibilities of those choices being dataset-specific and our goal of comparing 3 strategies on the same ground, batch size is set to 8 and training input frame number to 25 for all BasicVSR++ experiments, which proceed 200k iterations with all other hyperparameters set to default values. For the same reason, we kept parameters the same as their original denoising task for Restormer [15], except for progressive training scheme, which was changed to the schedule in Table 6. Based on its denoising settings, Uformer [13] was trained for 500 epochs with patch size being 256 and batch size 16 for all experiments. LEDNet [16] was trained with GAN structure for 200k iterations.

Reported in Table 7 are computational costs and processing speed of the models restoring full resolution (640×480) test frames.

Table 6. Progressive training schedule for Restormer [15]

Stage	1	2	3	4	5	6
Mini-batch size	8	5	2	1	1	1
Iteration number	46000	32000	24000	18000	18000	12000
Patch size	128	160	192	256	288	320

Table 7. Computational cost and speed for each model.

	FLOPs (G)	Params (M)	Speed (fps)
RVRT	2500.2	12.785	4.6764
BasicVSR++	1300.0	9.7602	31.374
Restormer	660.89	26.112	8.1294
Uformer	825.55	50.881	6.5128
LEDNet	180.88	7.4089	91.283

B. Theoretical Justification

In our main paper, we concluded that with an increase in the enhancement ratio, superiority of performance can shift from Gain-first to Exposure-first strategy. At ratio100, though in order to heat up the competition fairly our camera settings blur the lines between each strategy so that they all fall into a mixture of being enhanced by gain and exposure, it can be noticed that Mixed strategy has relatively better and more robust performances than the other two. We intend to offer insights for these phenomena from the formation of noise and blurriness in this section.

B.1. Basics

Previous works [7, 10, 11] that delivered analysis on related topics, *i.e.*, HDR and denoising vs. deblurring, are inherently different from ours in the way that they either ignored complexity of low-light noises, formulating calculations based on simplified Gaussian noise, or failed to compare strategies on fair grounds, leaving deterioration induced by short exposure uncompensated by cameras' native gain. Therefore, we propose to attack this topic from a distinct view which should be much closer to our experiment settings than precedent analyses.

Starting from the image formation details for low-light frames. As much as we want to include every minutiae of digital camera circuit designs as described in [12], it will be a herculean task to carry the weight of all those elements for further analysis. In the meantime, recent works [3, 6, 14] provide simplified models for noises generated by a camera from pragmatic viewpoints without loss of much details for enhancement tasks. Models proposed by Feng *et al*. [6] and Cao *et al*. [3] are similar to the one by Wei *et al*. [14] but merely with more accurate calibration and data augmentation techniques. Thus, for simplicity, we borrow notations from the latter for following derivations.

According to [14], acquisition of a digital image D can be formulated as follows,

$$D = ISP(KI + N), \tag{1}$$

where I is the number of photoelectrons induced by incident light, K being overall system gain, N being noise components, and ISP stands for image signal processing that transforms raw sensor data into sRGB domain with demosaicing, gamma correction, white-balancing, etc. The noise N mainly contains 4 parts, as shown in Equation 2: photon shot noise N_P that, combined with I, follows a Poisson

distribution due to the wave-particle duality of photons, *i.e.*, $(I + N_P) \sim \mathcal{P}(I)$; read noise N_{TL} that can be modeled by a Tukey lambda distribution, $N_{TL} \sim TL(\lambda; 0, \sigma_{TL})$; row noise N_r which imposes bias on each row by a zero-mean Gaussian distribution, $N_r \sim \mathcal{N}(0, \sigma_r)$; and quantization noise N_q for rounding error introduced by ADC converter, $N_q \sim U(-1/2q, 1/2q)$.

$$N = KN_P + N_{TL} + N_r + N_q \tag{2}$$

The logarithm of parameters for N_{TL} and N_r is logarithmically proportional to K, *i.e.*, $log(\sigma_{TL}) \propto log(K)$, $log(\sigma_r) \propto log(K)$, with a Gaussian perturbation term of fixed standard deviation ($\bar{\sigma}_{TL}$, $\bar{\sigma}_r$) added to each of them separately. For more details, please see equation 12 in [14].

As for exposure, we establish our foundation from the work of Cao *et al.* [2], for their consideration of authentic blurry image synthesis. Ideally, raw signal of a blurry image B_{real} with an exposure time τ should be $B_{real} = \int_0^\tau I(t)\, dt$. Yet, as reconstructing temporal continuous light signals from discontinuous ground truths is not a trivial step, we adopt the discrete form for the formation as in Equation 3 to keep our model tractable.

$$B_{raw} = \frac{1}{M} \sum_{t=0}^{M} S(t) + N \tag{3}$$

In the above equation, M is the number of latent sharp frames during the exposure time τ. $S(t)$ is the latent sharp and bright frames at moment t, *i.e.*, raw ground truth images at high frame rates, which we will synthesize later. The relationship between S and I in Equation 1 can be written as $I(t) = S(t)/gain$, where camera $gain$ is in linear scale, not logarithmic. For Gain-first, $M = 1$, whereas for Exposure-first $M = ratio$ while $ratio$, referring to the environmental light intensity reduction ratio, varies between 1 and 100 in our analysis, covering 4 categories of experiment conditions in our main paper.

When combining Equation 2 and 3 together, we carefully examined the hidden premises for all the simplifications. Because N_{TL} is a unified term that absorbs dark current noise, thermal noise and source follower noise, which are all constantly present during an exposure time, this term should be amalgamated into the summation of sharp frames, and similarly, photon shot noise N_P as well. The rest of the noise, N_r, which is a representation of banding pattern noise, mainly comes from sensor readout, downstream amplification, and ADC [12], thus treated as a one-time factor rather than an accumulating term. Therefore, a complete formula for a digital image produced by our strategies can be written as:

$$D = ISP\left(\frac{1}{M}\sum_{t=0}^{M}\left[\left(\frac{S_{raw}(t)}{gain} + N_P\right)\cdot gain + N_{TL}\right] + N_r + N_q\right) \tag{4}$$

with $M \times gain \equiv ratio$ for $M \in [1, ratio]$ to cover Gain-first, Exposure-first, and all possible Mixed strategies.

B.2. Metric

Instead of measurements that directly depict how far an image deviates from GT, like PSNR, SSIM, and LPIPS, here we want to evaluate the potential of recovering target signals, an upper bound for deep learning algorithms' performances. Hence, seeking help from information theory, we find it suitable to evaluate mutual information [8] between degraded inputs and GTs. It can be defined as follows:

$$MI(D, S) = \sum_{y \in S}\sum_{x \in D} P_{(D,S)}(x, y) log\left(\frac{P_{(D,S)}(x, y)}{P_D(x)P_S(y)}\right) \tag{5}$$

Joint distribution $P_{(D,S)}(x, y)$, and marginal distributions $P_D(x)$, $P_S(y)$ can be acquired by 2D and 1D histograms on synthetic degraded input D and ground truth S, both of which are in sRGB format.

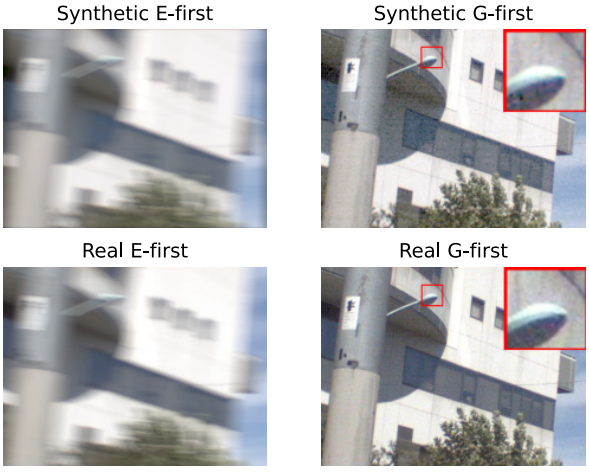

Figure 9. Authenticity of our synthetic images. Synthesized ratio30 Exposure-first blurry image can reach a PSNR of 33.37, while a synthetic Gain-first noisy image has a marginal KL-divergence of 0.03075 from the best parameter combination found by the grid search. As a frame of reference, test input PSNRs at this ratio is 26 $\sim$ 27 for G-first, 18 $\sim$ 26 for E-first, as reported in Table 8.

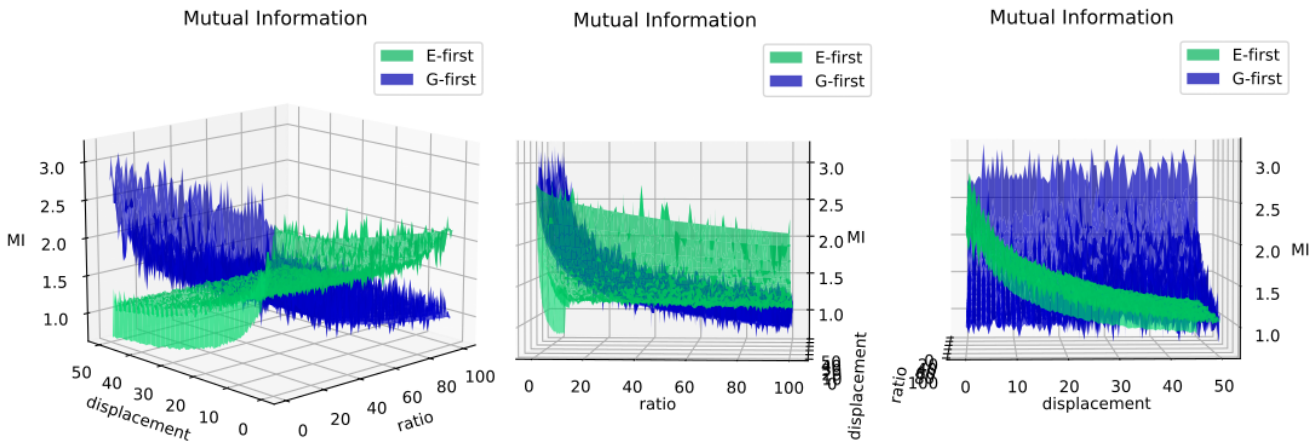

Figure 10. 3D plot of mutual information for G-first and E-first w.r.t. ratio and inter-frame pixel displacement. As can be seen from the pictures, increasing ratio and movement would impose damages on G-first and E-first respectively, showing separate areas where either one of them could be above the other. Better viewed in interactive 3D plot that can be generated from our submitted code.

B.3. Synthesis

Though it is possible to directly calculate mutual information of Gain-first images, such metric on Exposure-first and Mixed ones would depend largely on relations between frames, *i.e.*, distributions and correlations of $S_{raw}(t)$ across time among all pixels, which can easily become intractable when complicated scene or object motion exists. Therefore, to present evidence for our conclusions holding true in complicated real-world situations, we aim to prove our conclusions on a controllable synthetic movement, providing a simple instantiation without loss of generality.

Specifically, we translate a sharp image S horizontally and vertically through affine transformation matrix applied to each channel of RGGB separately to maintain correct color pattern in raw domain. Noise parameters at specific gain levels are calibrated by a grid search of possible values to minimize mean square error between generated images and image quadruplets in our dataset. For those parameters that capture essence of linearity between noises and K, as well as their uncertainty, *i.e.*, $\bar{\sigma}_{TL}$ and $\bar{\sigma}_r$, we use published results from [14] on a SonyA7S2 camera, which was then linearly stretched to fit data points from our camera found by the grid search. Utilizing this model, we are able to create samples that can have PSNR over 33 and marginal KL-divergence [1] below 0.04 when compared to captured real images, as shown in Figure 9, proving the validity of performing further analysis based on this synthesizing framework.

B.4. G-first vs E-first

In Equation 4, $M = 1$ corresponds to Gain-first, while $M = ratio$ corresponding to Exposure-first. We vary $ratio$ from 1 to 100, and in the meantime change inter-frame displacements that are used to control speed of artificial movements in long-exposure images from 0 to 50 pixels. Resulting mutual information for both strategies can be viewed in Figure 10. An interactive 3D plot is available on our Project page to provide better sight for their geometry.

From the figure, it can be seen that increasing ratio could wreak havoc on G-first images, as its mutual information dives below E-first when ratio goes above 50 (middle one in Figure 10). Viewed from other angles, it can be observed that E-first is sensitive to both ratio and displacement, defeating G-first at relatively high ratios with low speed motion, which is in accordance with intuition that long exposure eliminates noises for still scenes under extreme dark environments.

To help better apprehend differences between the two, we plot in Figure 11 a heatmap of E-first's mutual information minus G-first's. Color gradient and contours on this difference map clearly show existence of turning points, or in this case an equilibrium line, where the amount of latent information that can be used to restore images in two strategies are in close proximity to each other. This also means that neither Gain-first nor Exposure-first can be an once-and-for-all method, since on different side of the equilibrium line performances of the two switch places, further enhancing the value of empirical results presented in our main paper. Further, it can be verified from the figure that the equilibrium line intersects ratio10, ratio30, and ratio100 roughly at 7, 20 and 40 pixels of inter-frame displacement, illustrating sensitivity to motion of E-first when compared to G-first, validating our conclusions in the main paper that Gain-first takes a leading role at ratio10 and ratio30 [1], de-

[1]Frame rate at ratio30 drops to 15fps, increasing the likelihood for inter-frame displacement to go beyond turning point value.

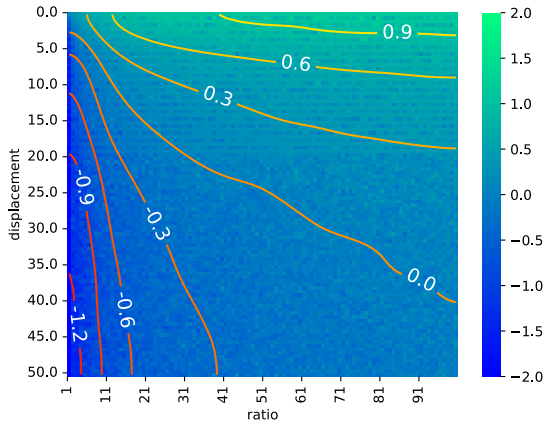

Figure 11. Difference between mutual information of E-first and G-first. Positive values indicate advantages for E-first, and negative for G-first. Contours are plotted on Gaussian filtered differences, for better visualization.

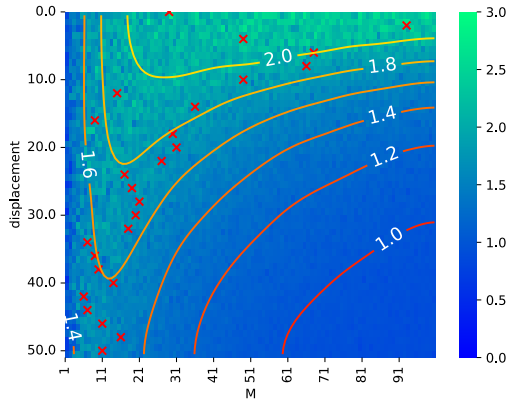

Figure 12. Mutual information of all possible strategies at ratio100. Red crosses indicate an optimal value M at each level of inter-frame motion. Contours are plotted on Gaussian filtered values, for better visualization.

clining at ratio100.

B.5. Mixed

To evaluate all possible combinations of gain and exposure, we focus on the extreme case, *i.e.*, ratio100, altering M and $gain$ accordingly to explore their effect on mutual information. Specifically, we traverse the integer value from 1 to 100 for M, setting $gain = \frac{ratio}{M} = \frac{100}{M}$ at the same time. An interactive 3D plot can be generated by our code, whereas here we present its heatmap with contours in Figure 12. We use red crosses to display an optimal value of M in that row, namely a best version of Mixed strategy at that level of displacement. It can be seen that such optimal solutions are clustered in area close to Exposure-first when displacement is small, and biased towards Gain-first with moving speed increasing. Yet, it never reaches to a point where it reduces to completely Gain-first ($M = 1$) or Exposure-first ($M = 100$), corroborating our conclusion that under this ratio a joint-denoising-deblurring process on the Mixed strategy images is preferred.

Table 8. Metrics of test inputs, evaluating the starting point for each experiment.

PSNR/SSIM		G-first	Mixed	E-first
Ratio10 15fps	Overall	30.56/0.8046	30.08/0.8836	26.98/0.8490
	Static	30.02/0.8002	32.37/0.9119	31.64/0.9174
	Moving	31.10/0.8091	27.79/0.8553	22.32/0.7807
Ratio10 30fps	Overall	30.56/0.8032	29.91/0.8837	26.30/0.8430
	Static	29.85/0.7923	31.98/0.9046	30.61/0.9040
	Moving	31.27/0.8141	27.84/0.8628	21.99/0.7820
Ratio30 15fps	Overall	26.70/0.6920	29.12/0.8499	22.59/0.7704
	Static	26.16/0.6759	30.53/0.8755	26.99/0.8847
	Moving	27.25/0.7080	27.71/0.8243	18.20/0.6562
Ratio100 30fps	Overall	22.26/0.4322	27.50/0.6832	26.52/0.7630
	Static	22.53/0.4512	27.36/0.6886	25.78/0.7467
	Moving	21.99/0.4133	27.64/0.6778	27.25/0.7793

Table 9. Averaged metric increase for each experiment. They are differences between metrics on test inputs as in Table 8 and metrics on restored results averaged among 5 enhancement algorithms we chose for each ratio and frame rate category.

PSNR/SSIM		G-first	Mixed	E-first
Ratio10 15fps	Overall	5.828/0.1543	5.122/0.0672	5.248/0.0675
	Static	6.134/0.1571	4.080/0.0453	3.430/0.0286
	Moving	5.520/0.1514	6.164/0.0891	7.066/0.1064
Ratio10 30fps	Overall	6.190/0.1569	5.690/0.0692	5.988/0.0805
	Static	6.200/0.1635	4.796/0.0522	3.998/0.0380
	Moving	6.178/0.1502	6.580/0.0861	7.980/0.1229
Ratio30 15fps	Overall	6.726/0.2538	4.604/0.1001	3.562/0.0617
	Static	5.702/0.2610	3.272/0.0796	2.250/0.0229
	Moving	7.738/0.2467	5.936/0.1206	4.868/0.1004
Ratio100 30fps	Overall	9.434/0.4851	5.000/0.2487	4.568/0.1410
	Static	8.794/0.4704	5.144/0.2542	7.336/0.2033
	Moving	10.08/0.4995	4.856/0.2432	1.812/0.0786

C. Additional Evaluations and Visualizations

C.1. Visualizing Quantitative Measurements

We include a visualization of all quantitative results from our experiments in Figure 13 to better aid the understanding of our conclusions.

C.2. Relative Metric Increase

In order to assess relative improvements in each strategy, we calculated PSNR/SSIM between test inputs and their GTs as reported in Table 8. Numbers in the table are in accordance with the derivation we had in section B.4. We can have their relative metric increases as shown in Ta-

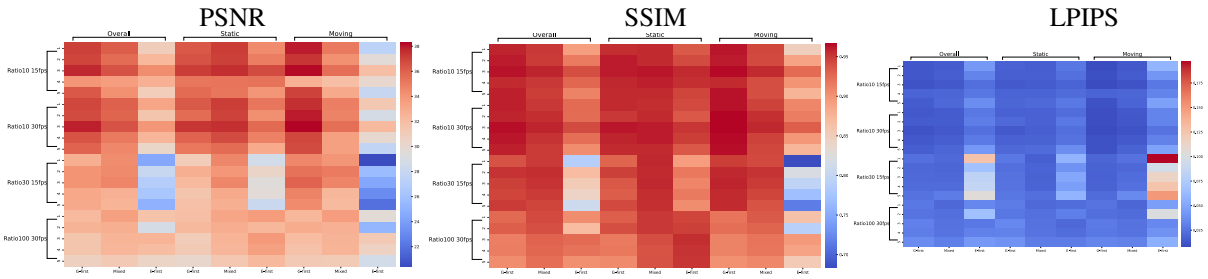

Figure 13. Visualization of quantitative measurements. Number on the row labels denote 5 methods we chose, *i.e.*, RVRT, BasicVSR++, Restormer, Uformer, and LEDNet, under 4 experiment conditions: ratio10 15fps, ratio10 30fps, ratio30 15fps, and ratio100 30fps. Columns include overll, static-shot, and moving-shot results of Gain-first, Mixed, and Exposure-first, integrating information from Table 3-5 in the main paper.

ble 9 by comparing these values to restored result measurements averaged among all five models, namely RVRT [9], BasicVSR++ [4, 5], Restormer [15], Uformer [13], LEDNet [16], for each experiment [2]. It can be seen that Gain-first has the most overall metric increase for all ratios, while Exposure-first and Mixed have larger boost on moving shots from ratio10 experiments.

D. Project Website Contents

We include more video samples for our collected dataset and experiment results on the Project page, where the dataset is publicly available. Other contents include a summary video and interactive 3D plots.

References

[1] Abdelrahman Abdelhamed, Marcus A Brubaker, and Michael S Brown. Noise Flow: Noise Modeling with Conditional Normalizing Flows. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[2] Mingdeng Cao, Zhihang Zhong, Yanbo Fan, Jiahao Wang, Yong Zhang, Jue Wang, Yujiu Yang, and Yinqiang Zheng. Towards real-world video deblurring by exploring blur formation process, 2022. 2

[3] Yue Cao, Ming Liu, Shuai Liu, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Physics-guided iso-dependent sensor noise modeling for extreme low-light photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5744–5753, 2023. 1

[4] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 5

[5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. On the generalization of BasicVSR++ to video deblurring and denoising. *arXiv preprint arXiv:2204.05308*, 2022. 1, 5

[2] Averaged measurements can be calculated from Table 2, 3, and 4 of our main paper

[6] Hansen Feng, Lizhi Wang, Yuzhi Wang, and Hua Huang. Learnability enhancement for low-light raw denoising: Where paired real data meets noise modeling. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 1436–1444, 2022. 1

[7] Samuel W. Hasinoff, Frédo Durand, and William T. Freeman. Noise-optimal capture for high dynamic range photography. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 553–560, 2010. 1

[8] Kisha Johnson, Arlene Cole-Rhodes, Ilya Zavorin, and Jacqueline Le Moigne. Mutual information as a similarity measure for remote sensing image registration. In *Geo-Spatial Image and Data Exploitation II*, pages 51 – 61. International Society for Optics and Photonics, SPIE, 2001. 2

[9] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. *arXiv preprint arXiv:2206.02146*, 2022. 1, 5

[10] Kaushik Mitra, Oliver Cossairt, and Ashok Veeraraghavan. To denoise or deblur: parameter optimization for imaging systems. In *Digital Photography X*, page 90230G. International Society for Optics and Photonics, SPIE, 2014. 1

[11] Kaushik Mitra, Oliver S. Cossairt, and Ashok Veeraraghavan. A framework for analysis of computational imaging systems: Role of signal prior, sensor noise and multiplexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):1909–1921, 2014. 1

[12] Junichi Nakamura. *Image Sensors and Signal Processing for Digital Still Cameras*. CRC Press, Inc., USA, 2005. 1, 2

[13] 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 (CVPR)*, pages 17683–17693, 2022. 1, 5

[14] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8520–8537, 2022. 1, 2, 3

[15] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang.

Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 1, 5

[16] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *ECCV*, 2022. 1, 5