

Training Neural Networks on RAW and HDR Images for Restoration Tasks

Andrew Yanzhe Ke¹ Lei Luo² Xiaoyu Xiang² Yuchen Fan²
Rakesh Ranjan² Alexandre Chapiro² Rafał K. Mantiuk¹
¹University of Cambridge ²Meta

Abstract

The vast majority of standard image and video content available online is represented in display-encoded color spaces, in which pixel values are conveniently scaled to a limited range (0–1) and the color distribution is approximately perceptually uniform. In contrast, both camera RAW and high dynamic range (HDR) images are often represented in linear color spaces, in which color values are linearly related to colorimetric quantities of light. While training on commonly available display-encoded images is a well-established practice, there is no consensus on how neural networks should be trained for tasks on RAW and HDR images in linear color spaces. In this work, we test several approaches on three popular image restoration applications: denoising, deblurring, and single-image super-resolution. We examine whether HDR/RAW images need to be display-encoded using popular transfer functions (PQ, PU21, and mu-law), or whether it is better to train in linear color spaces, but use loss functions that correct for perceptual non-uniformity. Our results indicate that neural networks train significantly better on HDR and RAW images represented in display-encoded color spaces, which offer better perceptual uniformity than linear spaces. This small change to the training strategy can bring a very substantial gain in performance, between 2 and 9 dB.

1. Introduction

Neural networks are typically trained on images represented in *display-encoded* color spaces, such as BT.709 with the sRGB non-linearity [2]. These are known as *display-encoded* color spaces because they are commonly used to drive displays. HDR images, as well as RAW images containing camera sensor values, are often represented in *linear color spaces*, in which the pixel values are proportional to colorimetric, photometric or radiometric light quantities¹. Notably, linear color spaces are very far from being percep-

tually uniform [17]: the same change in physical units is much less visible at high brightness than low. Modern perceptually uniform transfer functions, such as the SMPTE PQ (perceptual quantizer) [21] and PU21 (perceptually uniform transform) [18] have been developed to address this. By way of example, according to the PQ function, a change of 1 cd/m² starting from near darkness (0.005 cd/m²) is over 150 times more visible than the same change starting at 100 cd/m².

There are arguments both for and against training neural networks with images in linear spaces. If a plain L1 or L2 loss is used to train networks for linear data, we run the risk that the network will overfit for bright colors, and introduce unacceptable inaccuracies for darker colors. Additionally, if we wish to avoid visible quantization errors, pixel values in linear color spaces must be represented with more bits than display-encoded spaces. However, any physical phenomena involving light, such as lens blur, motion blur, or noise, can only be modeled in a physically plausible manner in linear color spaces. Therefore, physically plausible image formation models for blur, noise and sampling must be defined in linear color spaces, and by analogy, the inverse problems of deblurring, denoising and super-resolution should also be formulated in linear color spaces. It is unknown, however, whether neural networks can benefit from this connection to physical properties of linear color spaces.

In this work, we want to determine what training strategy leads to the best results. Should we train neural networks in display-encoded or linear color spaces (Section 3)? If we train networks in linear color spaces, should we use a loss function that compensates for perceptual non-uniformity (Section 4)? There is no consensus in the literature on these questions. Many works train on images in linear color spaces and use an L1 loss [6, 7, 28]. Some use a loss with perceptual encoding [8, 13, 20]. Our goal is to find empirical evidence to motivate a training strategy for image restoration techniques trained on HDR/RAW images. To that end, we collect datasets of HDR and RAW images and train six networks to perform common image restoration tasks: single-image super-resolution, denoising, and deblurring (Section 5).

¹The proportionality is approximate, as camera spectral sensitivity does not allow for direct measurement of colorimetric or photometric quantities.

Rather than proposing a new method, this paper is a benchmark and reproduction study intended to settle these open questions. Our results serve as best practice guidelines for future ML methods operating on HDR and RAW content. Our results (Section 6) show a clear benefit in training networks in display-encoded color spaces, offering a gain in performance between 2 and 9 dB for these applications.

2. Related work

2.1. Imaging RAW pixels

RAW image data, captured directly from image sensors, preserves unprocessed and uncompressed pixel data, thereby providing rich information for restoration techniques. RAW image formats also represent pixels in linear color spaces, which facilitate physically plausible modeling of camera noise [1]. Early works [10, 12] in this area mainly focused on noise reduction and artifact suppression based on statistical models of noise. With the advent of deep learning, restoration in the RAW domain has witnessed a paradigm shift. [4, 22, 26] pioneered by the integration of deep neural networks for denoising. Those early works, however, targeted sRGB images, typically with synthetic additive Gaussian noise. Since denoising is likely to be integrated in the very early stages of the ISP pipeline, it should operate on linear RAW pixel values instead.

Surprisingly, many deep-learning methods proposed for processing RAW images were never tested on RAW images [16, 27]. Instead, they opted to work with display-encoded images in the sRGB space, and assumed additive Gaussian noise. Both assumptions make input data very different from RAW sensor pixel values, raising the question on whether these methods would be equally effective for real camera data. Other methods that were trained and tested on RAW images [6, 28], did so directly on linear color values and used an L1 loss. Our work demonstrates this strategy often fails to achieve good results.

2.2. High Dynamic Range (HDR) imaging

The dynamic range is the contrast between the brightest and darkest parts of an image or video. Notably, the dynamic range the human eye can see is extremely high and much larger than what most standard dynamic range (SDR) capture and display pipelines are capable of. High Dynamic Range (HDR) is a term for a range of technologies that enhance the visual quality and realism of content by providing a wider than normal dynamic range [23].

A key problem in HDR imaging involves merging multiple exposures into an HDR image while reducing “ghosting” artifacts due to motion [13]. Another problem is tone-mapping of HDR scenes to the dynamic range that can be shown on a display [23]. Yet another is the reconstruction of an HDR image from a single exposure (i.e. inverse-tone

mapping) [8, 14]. However, standard reconstruction problems like denoising, single-image super-resolution, or deconvolution/deblurring, are also relevant for HDR imaging.

This work focuses on how HDR data in linear color spaces was handled in these previous works. Both Eilertsen et al. [8] and Kalantari et al. [13] trained the networks to directly predict HDR images in a linear color space. However, to compensate for the perceptual non-uniformity of that space, Eilertsen et al. used the logarithmic function, and Kalantari et al. introduced the μ -law (explained in detail in Section 3). Mildenhall et al. [20] used a relative loss, similar to SMAPE (explained in Section 4), to train a NeRF on RAW images. The networks intended to reconstruct HDR UHD video content [14] were directly trained to predict display-encoded frames using the PQ OETF [21].

3. Representations

In this work, we explore popular perceptual pixel value representations from the literature for use in machine learning applications. The main question that arises is what representation — linear or display encoded — is most suitable for deep-learning networks. The linear representation should be better to model physical phenomena (assuming neural networks are able to take advantage of this benefit). However, the display-encoded representations are more data-efficient (requiring fewer bits) and have better perceptual uniformity. Moreover, most existing methods have been trained on display-encoded representations, as these are commonly used for SDR content widely available online. To answer the question above, we tested four representations: one linear and three display-encoded:

Linear: linear RGB pixel values with BT.709 primaries, scaled between 0 and 1;

μ -law: a transfer function commonly used for perceptual audio encoding. Kalantari and Ramamoorthi [13] proposed to encode HDR images using the μ -law function:

$$v_{\mu}(l) = \frac{\log(1 + \mu l)}{\log(1 + \mu)} \quad (1)$$

with the constant $\mu = 5000$. l is a relative linear (RGB) color value scaled to the range 0–1.

PQ (perceptual quantizer): display-encoding for HDR content [21], used in video coding and display standards (e.g., BT 2100). The PQ encoding was applied separately to each RGB channel in our experiments. Input linear values were given in absolute units (see Section 5.2).

PU21 (perceptually uniform transform): a transform similar to PQ but based on more modern data. It was originally proposed as an encoding of HDR images to be used with existing standard dynamic range metrics [3, 18]. For the PU21 encoding we use the quadratic function fitted to the original curve (which was derived numerically)

$$v_{\text{PU21}}(L) = a (\log_2(L) - L_{\min})^2 + b (\log_2(L) - L_{\min}), \quad (2)$$

where L is an absolute linear (RGB) color value (in the range 0.005–10 000) and the fitted parameters are $a = 0.001908$, $b = 0.0078$, and $L_{\min} = \log_2(0.005)$. The encoded values are between 0 and 1. The inverse is given by:

$$v_{\text{PU21}}^{-1}(V) = 2^{\frac{2aL_{\min} - b + \sqrt{b^2 + 4aV}}{2a}}. \quad (3)$$

The functions above were fitted to the PU21 encoding for a banding model without the influence of glare (details in [18]) and are more computationally efficient than the PQ encoding or the original source code for PU21. This encoding was applied to each linear RGB channel in the same manner as PQ (see Section 5.2).

The pixel encodings described above are plotted in Figure 1. It can be observed that both PQ and PU21 share very similar shapes, but PQ does not reach 0 for the minimum encoded value (0.005). The μ -law has a much shallower shape for values below 1, which means these values are encoded with less precision. Finally, linear encoding over-emphasizes large pixel values by allocating a very steep slope to those (note that the luminance is plotted on a logarithmic scale).

4. Loss functions

In addition to the way pixels are encoded, the loss function is a key feature of training neural networks. Most image restoration problems can be effectively trained using L1 or L2 norms on pixel values as a loss function. However, this is not necessarily true when pixel values are represented in linear color spaces (e.g. RAW or HDR images), as they are perceptually non-uniform — a small error at low brightness is more visible than the same error at large brightness levels. This problem can be avoided by first converting the compared pixel values to a color space that is more perceptually uniform so that the loss function becomes

$$\ell = \|p(f_\phi(\mathbf{x})) - p(\mathbf{y})\|_1 \quad (4)$$

where $f_\phi(\cdot)$ is a network with the weights ϕ , \mathbf{x} is input, \mathbf{y} is a reference (label) and the $p(\cdot)$ is a perceptual encoding function such as μ -law, PQ or PU21 (all introduced in Section 3). One important consideration is that the values passed to the function $p(\cdot)$ must be within the allowed range (e.g., greater than 0.005 for PQ and PU21). This

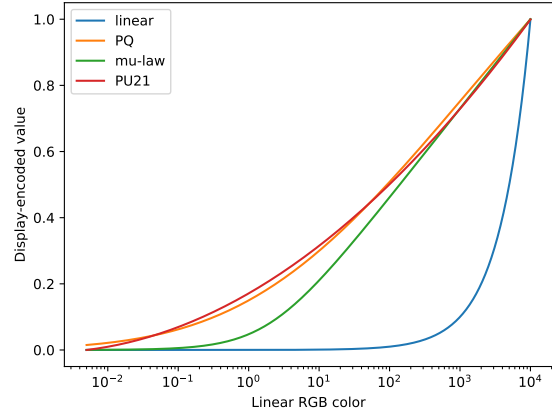


Figure 1. Perceptual encoding functions used to transform linear RGB color values to an approximately perceptually uniform space. A logarithmic scale is used for the x-axis to account for Weber’s law. Following BT.2100, the range of interest for encoded values is between 0.005 and 10 000. Note, however, that in our experiments, we used a smaller range between 0 and 4 000. Since μ -law encoding expects input values between 0 and 1, the values were divided by 10 000 before passing to the function in Eq. (1).

means that out-of-range values must be clamped. Both PQ and PU21 require absolute photometric values, which correspond to light emitted from a display (e.g., between 0.005 and 10 000).

The final loss we investigate is the symmetric mean absolute percentage error (SMAPE):

$$\ell = \left\| \frac{|f_\phi(\mathbf{x}) - \mathbf{y}|}{|f_\phi(\mathbf{x})| + |\mathbf{y}| + \epsilon} \right\|_1, \quad (5)$$

where ϵ is a small constant avoiding division by 0. Dividing by the sum of absolute values makes the error relative and more perceptually uniform, following Weber’s law.

5. Experiments

In this work, we focus on image restoration problems (denoising, deblurring, and single-image super-resolution), which aim to generate a higher-quality version of the input image. We exclude higher-level tasks (e.g., recognition) as they involve very different losses to those used for restoration. We also do not consider the tasks in which input and output images differ in dynamic range — tone mapping and single-image HDR reconstruction — as these tasks would only allow us to test either the representations (for tone mapping) or the loss functions (single-image HDR reconstruction) but not both at once.

To compare the effectiveness of different loss functions and pixel encodings, we train networks with HDR images on three tasks: image denoising, deblurring, and super-resolution. For each task, we choose two different neural

networks: one a well-established technique and the second a more recent method. We follow the training settings proposed in the original works. EDSR (2017) [15] and Real-ESRGAN (2021) [25] are used for the task of single-image super-resolution (4x). The downsampled images used for training were created using a bilinear filter. For image deblurring, we train GFNNNet (2018) [32] and MirNet-v2 (2022) [30]. The networks are trained on images affected by Gaussian blur with sigma values of 8px in both directions. DnCNN (2017) [31] and SADNet (2020) [5] are used for denoising of HDR images. All the tested methods were originally designed for display-encoded (SDR) images. Therefore, our tests demonstrate how well they generalize to HDR data and what is the best training strategy.

All networks are trained using the code provided by the authors on their GitHub repositories. We adopted the same training schemes as the original papers to train each network, including the optimizer and learning rate. The exception was Real-ESRGAN for which we used only L1 loss (disabled the perceptual loss) and adjusted the learning rates as listed in Table S1 (in supplementary). We trained each network for a different number of epochs to ensure that convergence was reached. EDSR was trained for 3k epochs, 150 epochs for Real-ESRGAN, 1k epochs for MirNet-v2, GFNNNet, DnCNN, and SADNet.

5.1. Datasets

We used two datasets: one for testing reconstruction on HDR images and another for testing on camera RAW images. The first dataset consisted of 106 HDR images from the HDR Photographic Survey [9] and HDR video frames from the SJTU HDR Video Sequences [24]. Since frames originating from the same video share similarities, we used one representative HDR frame provided for each video², adding an additional 16 reference images to the dataset. 60% of images were selected for training, 20% for validation, and the remaining 20% were used for testing. To verify the networks’ generalization ability to exposure variations, the testing data set was augmented five-fold by multiplying it with a random exposure coefficient drawn from a uniform distribution between 0.1 and 0.9. Physically plausible simulated camera noise (photon and readout noise) was added to training images [10].

Additionally, to show that our observations generalize to RAW images, we test super-resolution on the *Learning to See in the Dark* dataset [6]. We utilized 231 long-exposure RAW images from this dataset [6], with the same 60/20/20 split for training, validation, and testing. The color values were reconstructed from RAW values using Adaptive Homogeneity-Directed (AHD) demosaicing [11] and converted to the BT.709 color space (linear). The demosaicing

²The frames can be downloaded from the [website](#).

Table 1. The eight combinations of pixel encoding and loss functions tested in our experiments.

Label	Pixel encoding	Loss function
<i>Linear-L1</i>	Linear	L1
<i>PQ-L1</i>	PQ	L1
<i>PU21-L1</i>	PU21	L1
μ -L1	μ -law (mu-law)	L1
<i>Linear-PQ</i>	Linear	PQ
<i>Linear-PU21</i>	Linear	PU21
<i>Linear-μ</i>	Linear	μ -law (mu-law)
<i>Linear-SMAPE</i>	Linear	SMAPE

step was necessary as the tested architectures expected RGB images as input. Because RAW pixel values are relative, we had to rescale RGB values to the absolute range suitable for the PU21 and PQ encoding. We adjusted exposure of each image so that the average luminance was 20 nits. Such exposure adjustment matched the average luminance of SDR content and made images sufficiently bright without saturating the brightest pixels. The histograms of pixel values across both datasets can be found Figure S1 of the supplementary document. Note that due to the computational constraints, we could test only super-resolution methods on the RAW dataset.

5.2. HDR/RAW training cases

We tested 8 combinations of pixel encodings and loss functions, listed in Table 1. All perceptual pixel encodings (PQ, PU21, and μ -law, see Section 3) were used with a regular L1 loss because the encoded values are expected to be approximately perceptually uniform and, therefore, should not require a custom loss function. However, when no encoding was used (labeled *linear*), we used either an L1 loss or one of the losses described in Section 4. We also experimented with losses that were the sum of two terms, such as L1 and PQ-encoded L1, but they performed no better than individual loss terms. Figure 2 shows a diagram of our training setup.

Since PQ and PU21 both require pixel values in absolute units (of light, as emitted by a display), the linear pixel values were scaled to the range 0.005–4 000. Although both encodings allow values up to 10 000 (nit), 4 000 was selected as being more representative of realistic peak luminance values for a high-quality HDR display. For other pixel encodings, the linear color values were scaled in the 0–1 range. By following these steps, we ensure all values passed to the networks (for every encoding) were between 0 and 1, as is typical for training.

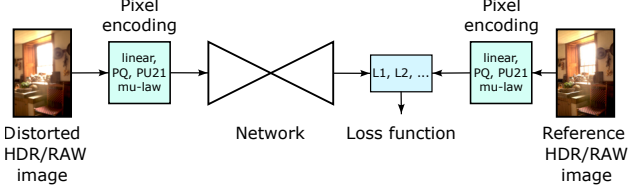


Figure 2. Training on HDR/RAW images. Training pairs were first (optionally) encoded from a linear to a perceptually uniform color space using PQ, PU21 or a μ -law transfer function (explained in Section 3). If no pixel encoding was applied (linear), one of the loss functions from Section 4 was used to account for the perceptual non-uniformity.

6. Results

To evaluate the results, we used three image quality metrics: PSNR, SSIM and ColorVideoVDP [19]. PSNR and SSIM were selected as the two most widely used metrics. ColorVideoVDP was selected because it natively supports HDR images and was designed to compare color images and videos. To adapt PSNR and SSIM to HDR images, we used the PU21 encoding [18], i.e. PSNR was computed on the PU21-encoded RGB pixels. SSIM was computed on the luma channel of the PU21-encoded pixels.

6.1. Single-image super-resolution

The numerical results for the two single-image super-resolution methods can be found in Figure 3 and in the top row of Table 2, and examples of visual results in Figure S2 (in supplementary) for HDR and in Figure S3 (in supplementary) for the RAW dataset. In Figure 3, we visualize the distribution of the quality scores across the entire test set and also report the results of pair-wise t-tests (two-tailed, $\alpha = 0.05$, $N = 144$ for HDR (24 test images \times 6 exposures) and $N = 276$ for RAW (46 test images \times 6 exposures)). The violin plots are sorted left-to-right from the best-performing condition to the worst. Red horizontal lines connecting two or more violin plots signify no evidence of statistically significant differences within that group. Conversely, conditions that are not joined by a common line were found to have performances that differ at a statistically significant level.

The results in Figure 3 indicate a substantial gain in performance of both super-resolution methods (1.5–9 dB, as compared to *Linear-L1*) when one of the perceptual pixel representations is used. However, we have no statistical evidence showing that one perceptual representation is better than the others. Interestingly, except for the ESDR-RAW experiment, perceptual loss functions did not improve the results of either super-resolution methods with respect to the “plain” *Linear-L1*. The visual results in Figure S3 and Figure S2 show that networks trained on *Linear* representations often result in color artifacts. The networks trained on perceptual representations (*PU-L1*, *PQ-L1* and μ -*L1*) do

not suffer from those artifacts and produce sharper images.

6.2. Deblurring

The deblurring results are different for the two tested methods and will be discussed separately. GFNNet failed to train and converge for several of the tested loss functions; it did, however, converge for all perceptual pixel encodings. As shown in Figure 4 and Table 2, the best performance was obtained using either the μ -law or PU21 pixel encodings, resulting in an improvement of about 3.8 dB compared to the plain *Linear-L1* configuration. The improvement was smaller (but still significant) for the PQ pixel encoding and linear encoding using the μ -law loss function. Visually (see Figure S4 in supplementary) the *Linear-L1* configuration resulted in artifacts in the form of vertical stripes. These artifacts were also present but less noticeable for other configurations. The *PQ-L1* configuration resulted in a slightly softer image.

Deblurring with MirNet-v2 produced better results than for GFNNet. Similarly as for GFNNet, the worst results were achieved by the plain *Linear-L1* configuration. The best results were obtained by a linear encoding with the μ -law loss, PQ and μ -law pixel encodings (gain of 2–3 dB). Interestingly, the PU21 pixel encoding resulted in worse results (see Figure S5 in supplementary, note the visual results’ softer appearance), though the reason is unclear.

6.3. Denoising

For denoising networks, PSNR had more distinctive and slightly different results than ColorVideoVDP and SSIM, as seen in Figure 5 and Table 2. We investigated this discrepancy further by comparing pairs of conditions for which the metric predictions differed the most. We observed that both denoising methods, in particular DnCNN, resulted in images that differed in brightness, tone, color, and contrast from the original. Even though such changes may be less perceptually objectionable (and difficult to notice without a reference), PSNR was very sensitive to them. Both ColorVideoVDP and SSIM put more emphasis on structural distortions and seemed to correlate better with the perceived quality.

For both denoising methods, perceptual pixel encodings produced the best results in most cases. μ -law and PU21 performed the best for DnCNN, followed by PQ, which performed slightly worse according to PSNR and ColorVideoVDP. In the case of SADNet, ColorVideoVDP could not distinguish between any of the configurations (no evidence of statistical differences). While μ -law consistently produced the highest metric scores for SADNet, PU21 was the second best in terms of SSIM, but the last in terms of PSNR. We attribute this discrepancy to color shifts (SSIM is insensitive to color changes). The visual results (see Figure S6 in supplementary) show changes in color for

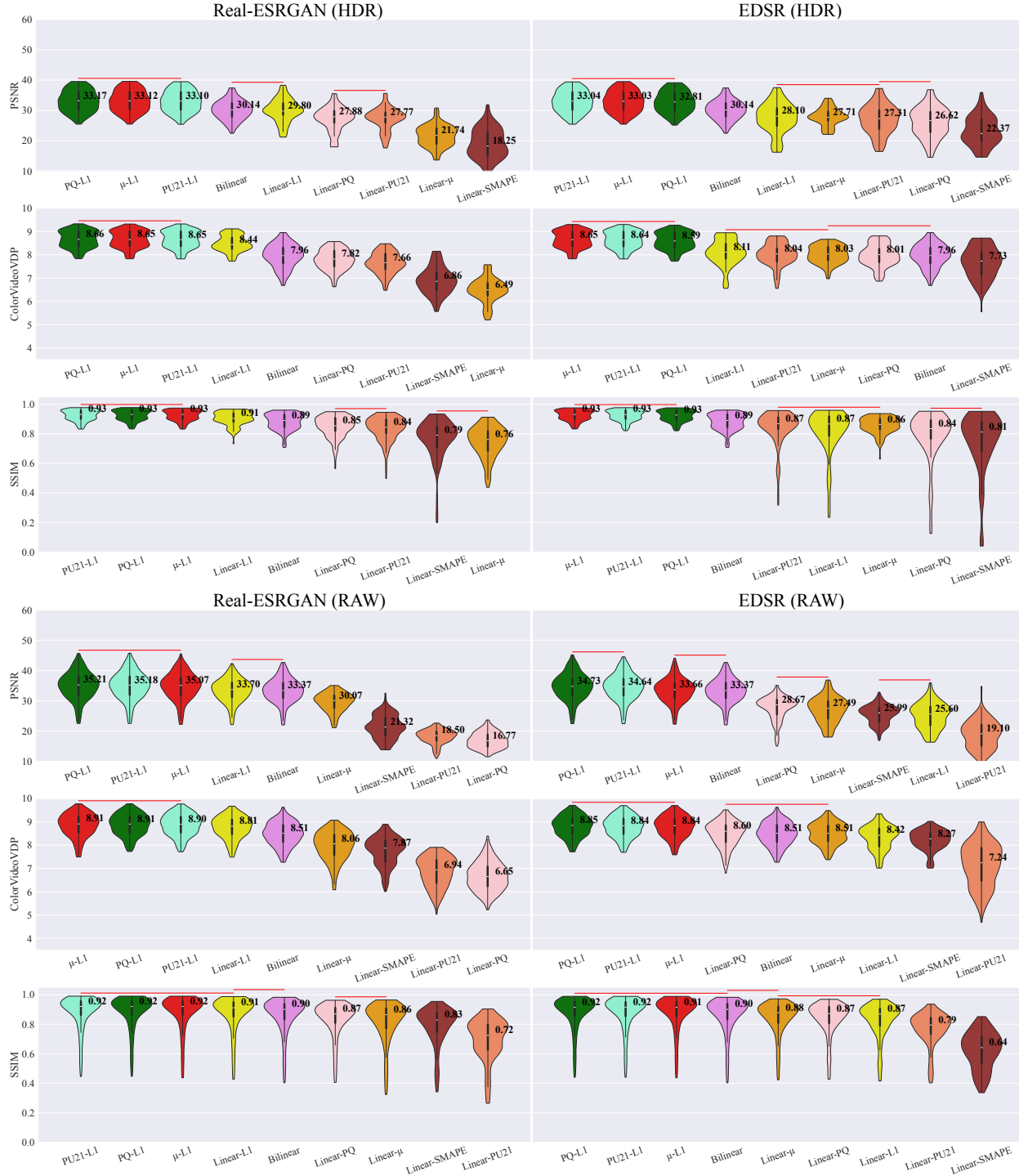


Figure 3. Single-image super-resolution results for the HDR (top) and RAW (bottom) datasets, the two networks (Real-ESRGAN and EDSR), and three metrics. The violin shape represents the distribution across the testing set. A thin black line shows the range of quality scores (excluding outliers), and the thicker line shows the region between the 25th and 75th percentiles. The dot and numerical value represent the median. Red horizontal lines above the violin plots denote the groups of conditions for which we have no evidence of statistically significant differences. Results are sorted from the best on the left to the worst on the right. The colors are different for each combination of representation and loss and are consistent across plots.

Table 2. The median values for evaluation results. The last column “Naive” is bilinear-upsampling results for super-resolution experiments, original blurry images for deblurring experiments, and the original noisy images for denoising experiments. The best-performing scores are annotated in red, while the second-best is in blue. The “N/A” entries correspond to the cases for which the training failed.

			PU21-L1	PQ-L1	μ -L1	Linear-L1	Linear-PU21	Linear-PQ	Linear- μ	Linear-SMAPE	Naive
SI-SR — HDR	EDSR	SSIM	0.93	0.93	0.93	0.87	0.87	0.84	0.86	0.81	0.89
		PSNR	33.04	32.81	33.03	28.10	27.31	26.62	27.71	22.37	30.14
		CVVDP	8.64	8.59	8.65	8.11	8.04	8.01	8.03	7.73	7.96
	Real-ESRGAN	SSIM	0.93	0.93	0.93	0.91	0.84	0.85	0.76	0.79	0.89
		PSNR	33.10	33.17	33.12	29.80	27.77	27.88	21.74	18.25	30.14
		CVVDP	8.65	8.66	8.65	8.44	7.66	7.82	6.49	6.86	7.96
	EDSR	SSIM	0.92	0.92	0.91	0.87	0.79	0.87	0.88	0.64	0.90
		PSNR	34.64	34.73	33.66	25.60	19.10	28.67	27.49	25.99	33.37
		CVVDP	8.84	8.85	8.84	8.42	7.24	8.60	8.51	8.27	8.51
SI-SR — RAW	EDSR	SSIM	0.92	0.92	0.92	0.91	0.72	0.87	0.86	0.83	0.90
		PSNR	35.18	35.21	35.07	33.70	18.50	16.77	30.07	21.32	33.37
		CVVDP	8.90	8.91	8.91	8.81	6.94	6.65	8.06	7.87	8.51
	Real-ESRGAN	SSIM	0.94	0.93	0.92	0.83	N/A	N/A	0.88	N/A	0.87
		PSNR	33.52	32.57	33.65	29.98	N/A	N/A	32.80	N/A	27.30
		CVVDP	7.59	7.27	7.66	7.09	N/A	N/A	7.41	N/A	5.51
	MirNet-v2	SSIM	0.93	0.95	0.94	0.93	0.93	0.95	0.95	0.95	0.87
		PSNR	32.99	34.65	34.10	31.98	32.06	33.57	34.80	34.10	27.30
		CVVDP	7.40	7.89	7.77	7.31	7.35	7.73	7.93	7.72	5.51
Deblurring	GFNet	SSIM	0.85	0.84	0.86	0.80	0.83	0.83	0.82	0.80	0.24
		PSNR	23.10	22.44	23.65	17.93	20.56	21.03	20.56	18.57	10.60
		CVVDP	6.07	5.96	6.15	5.68	5.87	5.74	5.76	5.66	4.37
	DnCNN	SSIM	0.90	0.91	0.90	0.88	0.89	0.89	0.88	0.89	0.24
		PSNR	23.94	27.76	26.47	24.62	26.03	25.32	26.32	26.08	10.60
		CVVDP	6.52	6.86	6.61	6.71	6.73	6.68	6.67	6.71	4.37
	SADNet	SSIM	0.85	0.84	0.86	0.80	0.83	0.83	0.82	0.80	0.24
		PSNR	23.10	22.44	23.65	17.93	20.56	21.03	20.56	18.57	10.60
		CVVDP	6.07	5.96	6.15	5.68	5.87	5.74	5.76	5.66	4.37
Denoising	DnCNN	SSIM	0.85	0.84	0.86	0.80	0.83	0.83	0.82	0.80	0.24
		PSNR	23.10	22.44	23.65	17.93	20.56	21.03	20.56	18.57	10.60
		CVVDP	6.07	5.96	6.15	5.68	5.87	5.74	5.76	5.66	4.37
	SADNet	SSIM	0.90	0.91	0.90	0.88	0.89	0.89	0.88	0.89	0.24
		PSNR	23.94	27.76	26.47	24.62	26.03	25.32	26.32	26.08	10.60
		CVVDP	6.52	6.86	6.61	6.71	6.73	6.68	6.67	6.71	4.37
	GFNet	SSIM	0.85	0.84	0.86	0.80	0.83	0.83	0.82	0.80	0.24
		PSNR	23.10	22.44	23.65	17.93	20.56	21.03	20.56	18.57	10.60
		CVVDP	6.07	5.96	6.15	5.68	5.87	5.74	5.76	5.66	4.37

Linear-PQ and Linear-PU. As a result, these two configurations have lower PSNR than the input noisy image (but higher ColourVideoVDP index because of the denoising).

7. Discussion and conclusions

The results across six tested methods clearly indicate that networks for image restoration should be trained on HDR/RAW images that are encoded using one of the perceptual transforms: PU21, PQ, or μ -law. We did not find sufficient evidence to recommend one transform over another across all applications. However, we observed slightly worse performance of PQ for GFNet, and PU21 for MirNet-v2 and SADNet. μ -law encoding was robust across different methods but did not always produce the best results. Both μ -law and PU21 use simpler and more computationally efficient formulas. PQ fails to scale the resulting values to a 0–1 range (see Figure 1), which may cause practical issues. Unlike the μ -law, PQ and PU21 were derived from psychophysical data and have stronger perceptual bases. As the differences in performance between the three transforms are mostly small, we recommend using any of them, as this offers a substantial gain in performance over not using them at all. The actual gain in performance is application-specific and can be as large as 2–9 dB as compared to the “default” approach of training linear color values with an L1 loss. Using linear color values in tandem with a loss function that employs perceptual

encoding seems to be less effective across the tested applications.

Although we cannot provide a conclusive answer of why perceptual transforms are beneficial for training neural networks, we suspect that this is because they offer better data efficiency. Pixels encoded with PU21, PQ, or μ -law require fewer bits to represent the same amount of visually relevant information. Although networks should be able to learn such a transform during training, this removes some capacity from the network, which could otherwise be devoted to their main task. Some works avoid the conundrum of what representation to use by providing both linear and gamma-encoded images to the network [13, 29]. This, however, leads to larger networks, reduced capacity or efficiency of the trained models.

The empirical findings in this work are relevant as they show that very small changes in training strategy (“one-liners”) can substantially improve the results of neural networks trained for popular tasks on HDR or RAW images. Moreover, the winning strategy of using perceptual pixel coding is rarely used in practice. More often, networks are trained in linear color spaces using an L1 cost function [6, 7, 28], or at most, use a perceptual encoding loss [8, 13, 20] — both strategies shown in this work to underperform. We hope our findings help practitioners in designing better training strategies for HDR/RAW images.

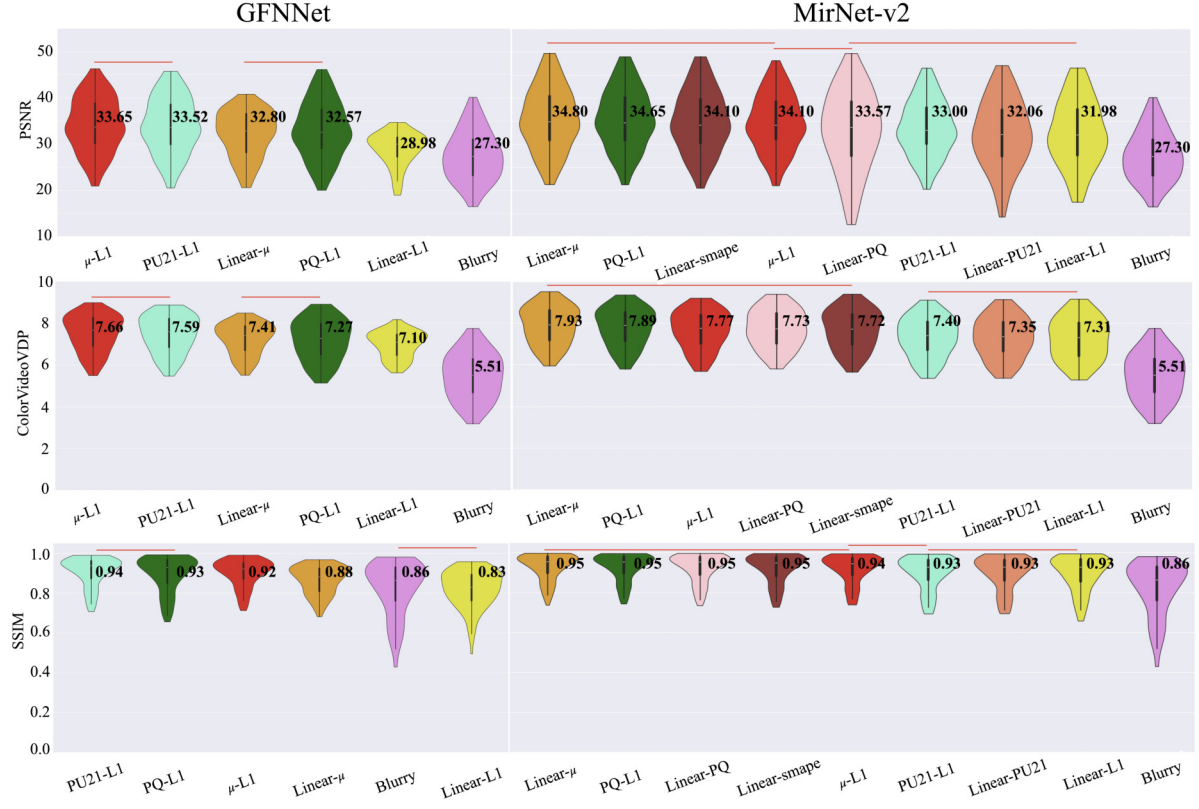


Figure 4. Deblurring results for the two networks (columns) and three metrics (rows). A few configurations are missing for GFNet as network training failed to converge for those. The notation is the same as in Figure 3.

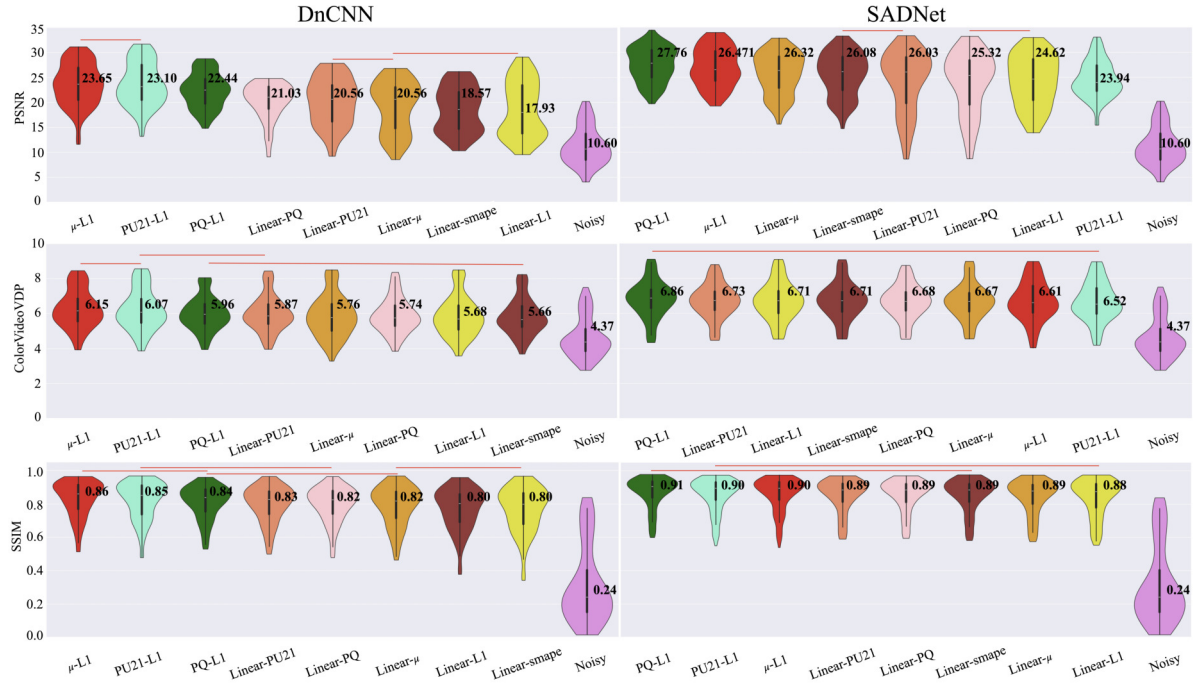


Figure 5. Denoising results for the two networks (columns) and three metrics (rows) The notation is the same as in Figure 3.

References

- [1] Cecilia Aguerrebere, Julie Delon, Yann Gousseau, and Pablo Musé. *Study of the digital camera acquisition process and statistical modeling of the sensor raw data*. 2013. hal-00733538. [2](#)
- [2] Matthew Anderson, Ricardo Motta, Srinivasan Chandrasekar, and Michael Stokes. Proposal for a standard default color space for the internet - sRGB. In *Color and Imaging Conference*, pages 238–245. Society for Imaging Science and Technology, 1996. [1](#)
- [3] Tunç O. Aydın, Rafal Mantiuk, and Hans-Peter Seidel. Extending quality metrics to full luminance range images. In *Human Vision and Electronic Imaging*, page 68060B. Spie, 2008. [3](#)
- [4] Harold C Burger, Christian J Schuler, and Stefan Harmeling. Image denoising: Can plain neural networks compete with BM3D? In *CVPR*, pages 2392–2399. IEEE, 2012. [2](#)
- [5] Meng Chang, Qi Li, Huajun Feng, and Zhihai Xu. Spatial-adaptive network for single image denoising. In *Computer Vision – ECCV 2020*, pages 171–187. Springer, 2020. [4](#)
- [6] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, page 3291–3300, Salt Lake City, UT, 2018. IEEE. [1](#), [2](#), [4](#), [7](#)
- [7] Marcos V. Conde, Florin Vasluianu, and Radu Timofte. Toward Efficient Deep Blind Raw Image Restoration. pages 1725–1731, Abu Dhabi, United Arab Emirates, 2024. IEEE. [1](#), [7](#)
- [8] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K. Mantiuk, and Jonas Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM Transaction on Graphics*, 36(6):Article 178, 2017. [1](#), [2](#), [7](#)
- [9] Mark D. Fairchild. The HDR Photographic Survey. *Color and Imaging Conference*, 15(1):233–238, 2007. [4](#)
- [10] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–54, 2008. [2](#), [4](#)
- [11] K. Hirakawa and T.W. Parks. Adaptive homogeneity-directed demosaicing algorithm. *IEEE Transactions on Image Processing*, 14(3):360–369, 2005. [4](#)
- [12] Youngbae Hwang, Jun-Sik Kim, and In So Kweon. Difference-based image noise modeling using skellam distribution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1329–1341, 2011. [2](#)
- [13] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36(4):1–12, 2017. [1](#), [2](#), [7](#)
- [14] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4k UHD HDR applications. In *CVPR*, pages 3116–3125, 2019. [2](#)
- [15] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR workshops*, pages 136–144, 2017. [4](#), [2](#), [3](#)
- [16] Lin Liu, Xu Jia, Jianzhuang Liu, and Qi Tian. Joint Demosaicing and Denoising With Self Guidance. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2237–2246. IEEE, 2020. [2](#)
- [17] Rafal Mantiuk, Grzegorz Krawczyk, Karol Myszkowski, and Hans-Peter Seidel. Perception-motivated high dynamic range video encoding. *ACM Transactions on Graphics (TOG)*, 23(3):733–741, 2004. [1](#)
- [18] Rafal K. Mantiuk and Maryam Azimi. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2021. [1](#), [3](#), [5](#)
- [19] Rafał K Mantiuk, Param Hanji, Maliha Ashraf, Yuta Asano, and Alexandre Chapiro. Colorvideovdp: A visual difference predictor for image, video and display distortions. *ACM Transaction on Graphics*, 43(4):129, 2024. [5](#)
- [20] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16169–16178. IEEE, 2022. [1](#), [2](#), [7](#)
- [21] S. Miller, M. Nezamabadi, and S. Daly. Perceptual Signal Coding for More Efficient Usage of Bit Codes. *SMPTE Motion Imaging Journal*, 122(4):52–59, 2013. [1](#), [2](#)
- [22] Seonghyeon Nam, Youngbae Hwang, Yasuyuki Matsushita, and Seon Joo Kim. A holistic approach to cross-channel image noise modeling and its application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1683–1691, 2016. [2](#)
- [23] Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski. *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010. [2](#)
- [24] Li Song, Yankai Liu, Xiaokang Yang, Guangtao Zhai, Rong Xie, and Wenjun Zhang. The SJTU HDR video sequence dataset. In *Proceedings of International Conference on Quality of Multimedia Experience (QoMEX 2016)*, page 100, 2016. [4](#)
- [25] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. [4](#), [2](#), [3](#)
- [26] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. *Advances in neural information processing systems*, 25, 2012. [2](#)
- [27] Wenzhu Xing and Karen Egiazarian. End-to-End Learning for Joint Image Demosaicing, Denoising and Super-Resolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3506–3515. IEEE, 2021. [2](#)
- [28] Xiangyu Xu, Yongrui Ma, Wenxiu Sun, and Ming-Hsuan Yang. Exploiting Raw Images for Real-Scene Super-Resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. [1](#), [2](#), [7](#)
- [29] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton Van Den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range

- imaging. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1751–1760, Long Beach, CA, USA, 2019. IEEE. 7
- [30] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):1934–1948, 2022. 4
- [31] 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. 4
- [32] Xinyi Zhang, Hang Dong, Zhe Hu, Wei-Sheng Lai, Fei Wang, and Ming-Hsuan Yang. Gated fusion network for joint image deblurring and super-resolution. *arXiv preprint arXiv:1807.10806*, 2018. 4

Training Neural Networks on RAW and HDR Images for Restoration Tasks

Supplementary Material

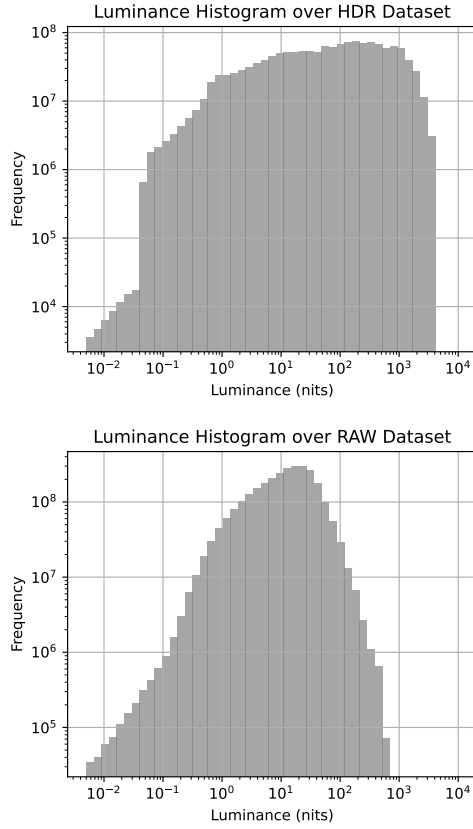


Figure S1. Histogram over all images in the HDR (top) and RAW (bottom) datasets.

Table S1. Learning rates used for Real-ESRGAN.

Label	Learning rate
Linear-L1	1×10^{-4}
PQ-L1	1×10^{-4}
PU21-L1	1×10^{-4}
μ -L1	1×10^{-4}
Linear-PQ	1×10^{-5}
Linear-PU21	1×10^{-5}
Linear- μ	1×10^{-5}
Linear-SMAPE	1×10^{-5}

S1. Content

This document contains additional results and details as listed below:

- Table S1 — learning rates used for Real-ESRGAN
- Figure S1 — histograms over all images and HDR and RAW datasets.
- Figure S2 — visual results for single-image super resolution on the HDR image dataset
- Figure S3 — visual results for single-image super resolution on the RAW image dataset
- Figure S4 — visual results for for deblurring with GFN-Net
- Figure S5 — visual results for for deblurring with MirNet-v2.
- Figure S6 — visual results for denoising — DnCNN and SADNet



Figure S2. Example results for single-image super-resolution on HDR images with EDSR [15] (top) and Real-ESRGAN [25] (bottom). The numbers in parentheses show PSNR, SSIM, and ColorVideoVDP quality values (the higher, the better) for the reconstructed image.

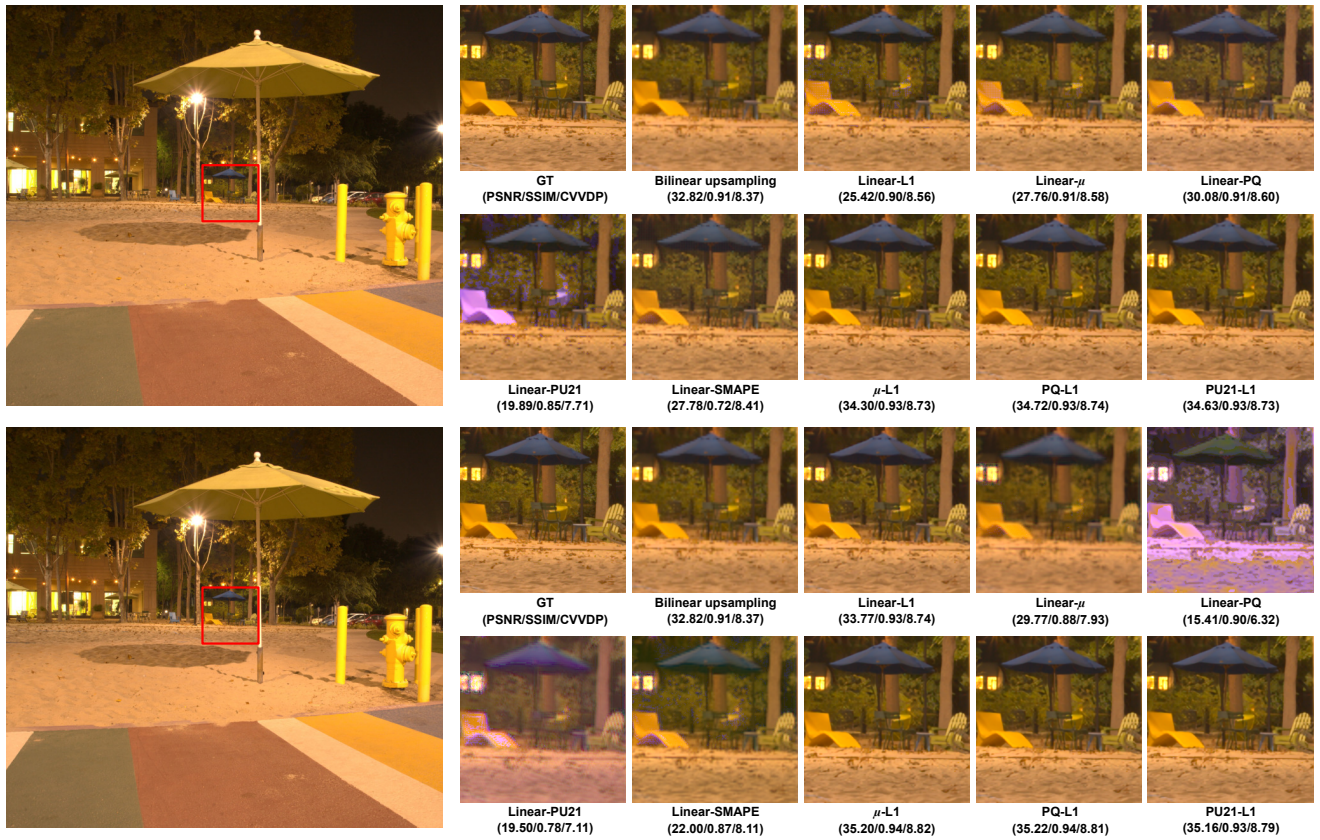


Figure S3. Example results for single-image super-resolution on RAW images with EDSR [15] (top) and Real-ESRGAN [25] (bottom).

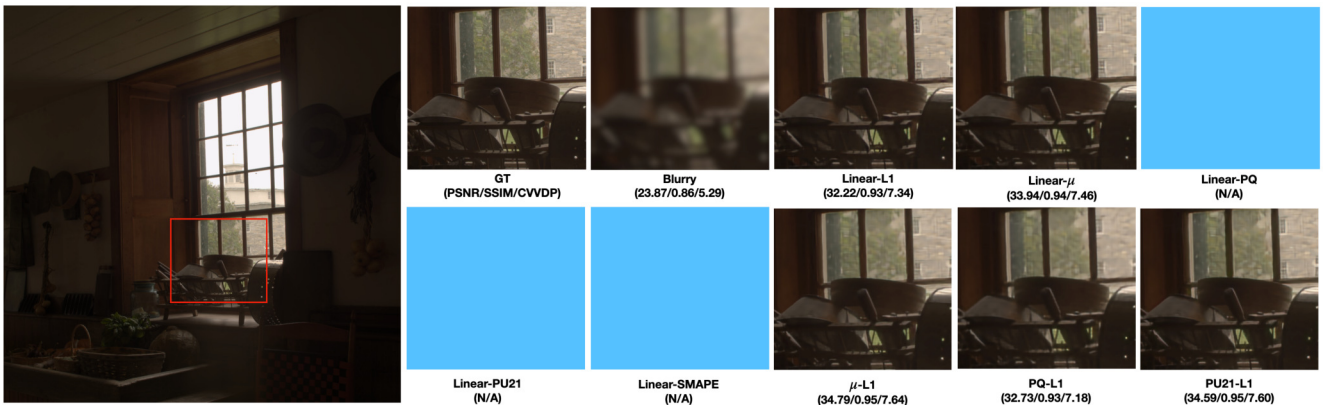


Figure S4. Example results for deblurring with GFNet. The results are missing for Linear-PQ, Linear-PU21, and Linear-SMAPE, as the networks failed to converge for these configurations.

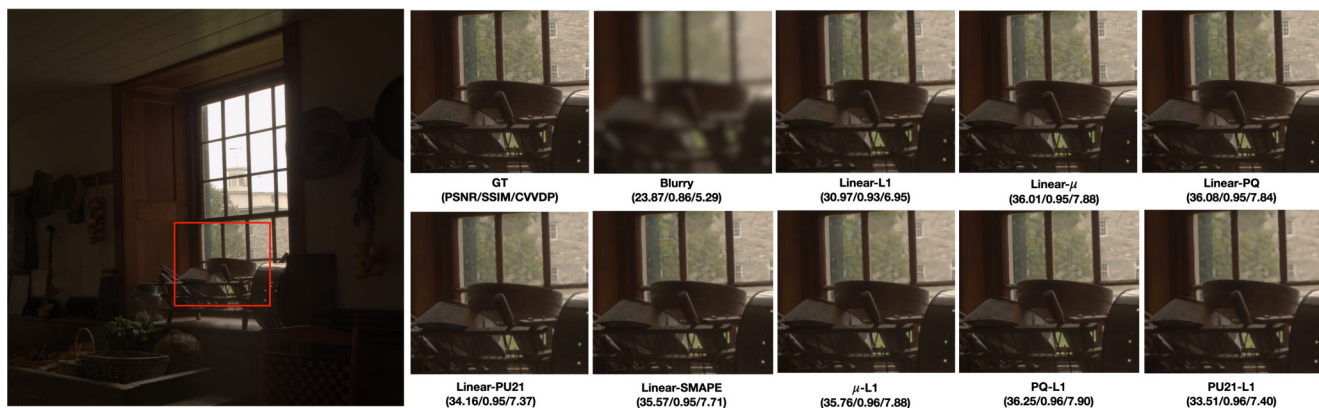


Figure S5. Example results for deblurring with MirNet-v2

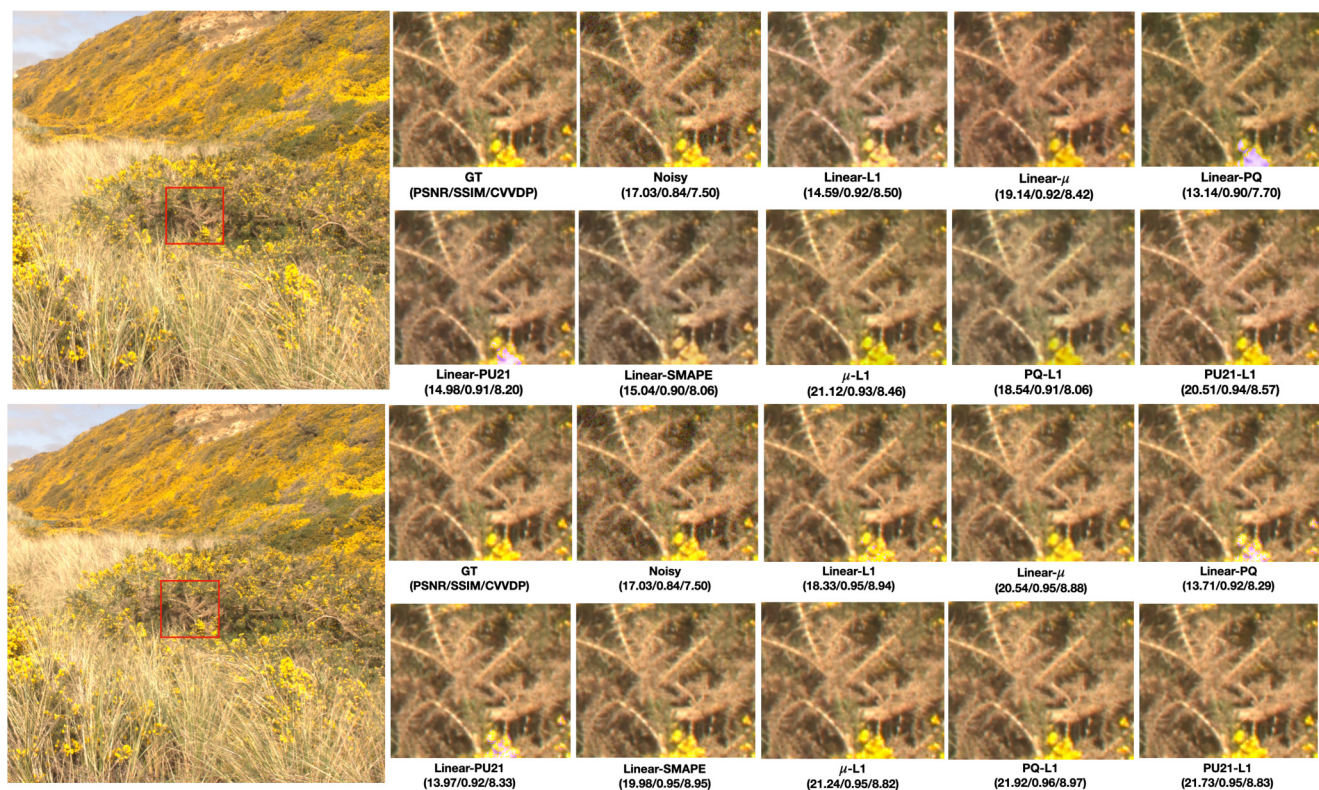


Figure S6. Example results for denoising with DnCNN (top) and SADNet (bottom).