The Devil is in the Dark Pixels: Toward Brightness Bias-Robust Denoising

Sungjun Cho[1], Zhuangzhuang Chen[1], and Xiaomeng Li[1⋆]

Department of Electronic and Computer Engineering,
The Hong Kong University of Science and Technology, Hong Kong SAR, China
schoaq@connect.ust.hk, eezzchen@ust.hk, eexmli@ust.hk

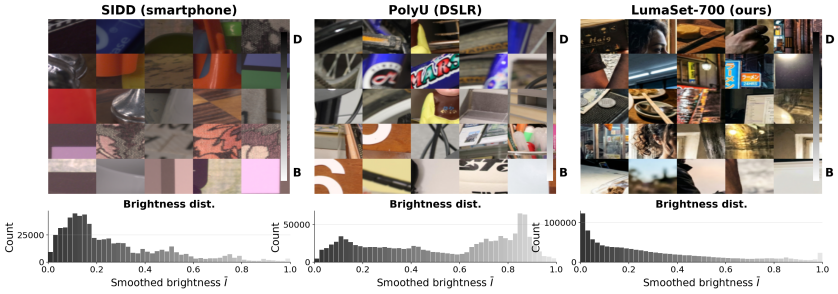
Abstract. In this paper, we reveal an important yet overlooked problem in image denoising: under signal-dependent camera noise models, dark regions suffer from inherently low Signal-to-Noise Ratio (SNR), as signal intensity decays far faster than noise variance diminishes, making detail recovery in dark areas fundamentally challenging. Yet rather than compensating for this difficulty, MSE-trained denoisers *exacerbate* it—reconstructing dark pixels up to 6× worse relative to their per-band noise floor. This bias stems from two compounding factors: signal-dependent noise inflates bright-pixel residuals, and the network's Jacobian norm increases monotonically with brightness. Together, these cause bright regions to chronically dominate gradient updates at the expense of dark ones. To this end, we propose *Brightness Bias-Robust Denoising* (BBRD), a drop-in replacement for MSE loss that partitions pixels into brightness bands, normalizes per-band error by empirical noise variance, and applies Group Distributionally Robust Optimization (Group-DRO) to dynamically upweight whichever band is currently worst, with zero additional parameters or inference cost. Across 8 architectures and 2 datasets in our experiments, BBRD is the **only method among 13 tested alternatives that improves each brightness band simultaneously**, achieving up to +0.45 dB on dark bands, +0.32 dB on bright bands, and +0.65 dB aggregate Peak Signal-to-Noise Ratio (PSNR) on SIDD, with the largest per-band gains in the darkest regions where detail recovery matters most. Code is available at `https://github.com/xmed-lab/BBRD`.

Keywords: Image denoising · Brightness bias · Signal-dependent noise · Distributionally robust optimization

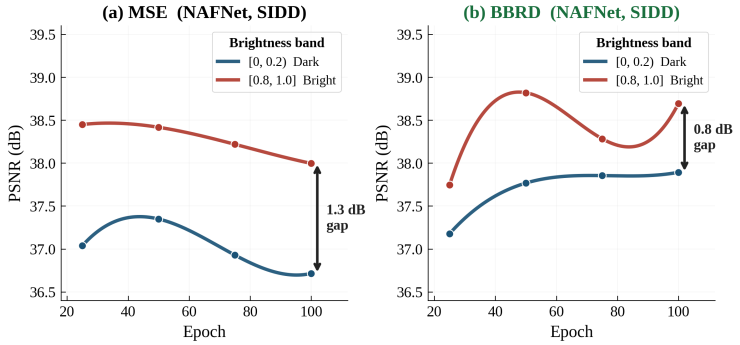
1 Introduction

Deep learning has driven rapid progress in image denoising [7, 13], with architectures advancing from CNNs [4, 34, 35] through Transformers [19, 32] to State-Space Models [9, 10], finding broad use in photography, medical imaging, and autonomous driving [5, 18, 23, 26]. Despite this architectural diversity, all models

⋆ Corresponding author.



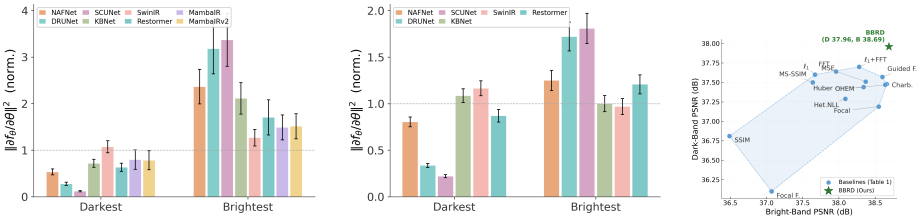
(a) Brightness distribution of SIDD, PolyU, and our dataset.



(b) Per-band PSNR during training (NAFNet, SIDD). Both panels use the same model, dataset, and brightness bands; only the loss differs (MSE vs. BBRD).

Fig. 1: (a) Brightness distribution across datasets. **(b)** Under identical NAFNet training on SIDD dataset, MSE under-optimizes dark regions (1.3 dB gap at convergence); BBRD reduces this to 0.8 dB with zero additional parameters.

share a common training paradigm: per-pixel MSE or ℓ_1 minimization, evaluated by a single aggregate PSNR. We reveal an important yet overlooked failure of this paradigm: *brightness bias*. Under a Poisson–Gaussian noise model, although absolute noise variance in dark regions is small, signal intensity decays far more rapidly, yielding inherently low Signal-to-Noise Ratio (SNR) in dark areas [22]. As shown in Fig. 1b, under MSE training the dark band decays to 36.7 dB while the bright band remains near 38.0 dB—a **1.3 dB gap that persists and even widens** as training progresses. At convergence, the darkest band of NAFNet [4] on SIDD achieves only 37.51 dB versus 38.37 dB for the brightest band, and exhibits 6× higher normalized error R_k relative to its per-band noise floor. This excess gap cannot be explained by SNR alone, pointing to a systematic failure in how these models are trained. This raises a fundamental question: ***Why do learned denoisers disproportionately under-optimize dark regions, compounding their already unfavorable SNR conditions?*** To answer this question, we conduct a series of experiments that train various denoising models on the SIDD dataset. As shown in Figs. 2a and 2b, the Jacobian norm of the brightest band (B5) consistently exceeds that of the darkest band



(a) Jacobian norm on SIDD **(b)** Jacobian norm on PolyU **(c)** SIDD

Fig. 2: Jacobian norm of the brightest band consistently exceeds that of the darkest band under MSE across 8 architectures (CNNs, Transformers, SSMs) on **(a)** SIDD and **(b)** PolyU, confirming the gradient imbalance is architecture-agnostic. **(c)** BBRD (PSNR-D 37.96, PSNR-B 38.69) strictly dominates all baselines—including the strongest, ℓ_1+FFT (37.70, 38.28) and MSE (37.51, 38.37)—pushing the Pareto frontier outward on SIDD with zero additional parameters.

(B1) across all 8 architectures spanning CNNs (NAFNet, DRUNet, SCUNet, KBNet), Transformers (SwinIR, Restormer), and State Space Models (MambaIR, MambaIRv2). This confirms that *brightness bias* is not attributable to model capacity or architectural inductive bias but to the intrinsic properties of the MSE loss function. The reason is twofold: (1) Under signal-dependent noise, brighter pixels naturally yield larger residuals, and (2) brighter pixels induce larger Jacobian norms. Intuitively, higher input activations produce larger intermediate feature magnitudes, which—after passing through any monotone or near-monotone nonlinearity (ReLU, GELU, SiLU, SimpleGate)—yield larger partial derivatives and thus amplify gradient flow. This is empirically confirmed across all 8 architectures despite their diverse activation choices (Figs. 2a and 2b). These two factors compound to dominate gradient updates, meaning the network's weights are updated primarily to fit bright regions. This gradient imbalance leaves dark regions chronically under-optimized (Sec. 3.2). This misalignment between what the loss optimizes and what the image actually needs reveals a fundamental limitation of uniform pixel-wise objectives in the presence of signal-dependent noise.

In light of this analysis, a straightforward solution is to adopt existing reweighting loss functions to focus on dark pixels. However, difficulty-based reweighting methods such as Focal Loss [20] and Online Hard Example Mining (OHEM) [25] mistakenly identify bright pixels as "hard" simply because they are signal-dependent, thus yielding larger residuals. As a result, these methods inadvertently *amplify* the bias by unintentionally doubling down on the existing gradient imbalance. Other objectives, e.g., robust pixel losses, uncertainty modeling, and frequency-domain losses, either operate uniformly over all pixels or fail to track which brightness band is currently lagging (comparison results can be found in Sec. 4.1). Moreover, there is no standard protocol to measure brightness disparity, and aggregate PSNR obscures the gap entirely, making the problem invisible to conventional evaluation.

To address these issues, we make the following contributions. First, we introduce a standardized brightness-disparity evaluation protocol for image denoising: per-band PSNR over five brightness strata, validated across 8 architectures and 13 loss functions. Second, we propose *Brightness Bias-Robust Denoising* (BBRD), a drop-in MSE replacement that corrects brightness bias via data-driven band partitioning, per-band noise normalization, and softmax Group DRO, with zero additional parameters or inference cost. Unlike existing losses that trade off between dark and bright performance, BBRD pushes the Pareto frontier outward. As shown in Fig. 2, in our experiments BBRD is the **only objective among 13 tested alternatives that improves every brightness band simultaneously**, achieving up to $+$**0.65 dB** aggregate PSNR on SIDD with visibly sharper dark-region textures (Fig. 5). Finally, as existing benchmarks lack balanced brightness coverage, we introduce **LumaSet-700**, a synthetic evaluation benchmark with controlled noise, and show that BBRD generalizes beyond signal-dependent noise. Our main contributions are summarized as follows:

- We introduce BBRD, a drop-in MSE replacement that corrects brightness bias via data-driven band partitioning, per-band noise normalization, and softmax Group DRO, achieving up to $+0.65$ dB on SIDD with zero additional parameters or inference cost (Sec. 3).
- We propose the first standardized per-band brightness-disparity evaluation protocol, demonstrating that brightness bias is universal across 8 architectures, 2 real-world datasets, and 13 loss functions (Sec. 4).
- We introduce **LumaSet-700**, a synthetic benchmark of 700 test patches stratified across dark, mid-tone, and bright scenes with controllable noise, demonstrating that BBRD generalizes beyond signal-dependent noise (Sec. 5.1).

2 Related Work

Deep image denoising. Deep image denoising has progressed through four major paradigms, each advancing architectural capacity while sharing a common training convention. CNN-based methods such as DnCNN [35], DRUNet [34], and NAFNet [4] established strong baselines using MSE or ℓ_1 loss. Transformer-based models including SwinIR [19] and Restormer [32] extended the receptive field via attention mechanisms, adopting Charbonnier loss for denoising. Diffusion models [6, 11, 21, 30] offer strong perceptual quality through iterative or distilled score-based refinement, but are trained by predicting added noise at each timestep, an objective equivalent to per-noise-level MSE with uniform pixel weighting. Most recently, State Space Models such as MambaIR [10] and MambaIRv2 [9] achieve linear-complexity global modeling, yet continue training with Charbonnier loss for denoising.

Notably, the above methods share the same critical blind spot: their base reconstruction objectives treat every pixel uniformly regardless of brightness, and are evaluated solely by aggregate PSNR, which renders brightness-stratified

disparities entirely invisible. More importantly, the dominant trend in image denoising research has been to improve model architecture—from CNNs to Transformers to State Space Models—while keeping the MSE loss fixed. BBRD takes the orthogonal direction: it keeps the model unchanged and corrects the loss, making it directly compatible with and validated on the latest SOTA architectures including MambaIR and MambaIRv2.

Loss functions for bias-robust learning. MSE (ℓ_2) remains the default for PSNR-oriented training. Alternative pixel-wise objectives like ℓ_1 and Charbonnier loss [3] are adopted in SwinIR [19] and Restormer [32] for sharper gradient behavior, while Huber loss [12] improves outlier robustness but does not account for structured noise heteroscedasticity. To address varying difficulty, Focal Loss [20] and OHEM [25] assign higher weights to harder examples; however, under signal-dependent noise, "hard" examples concentrate in high-variance bright regions, so these methods inadvertently *amplify* brightness bias rather than correcting it. Heteroscedastic NLL [16] models per-pixel aleatoric uncertainty by jointly learning $\sigma(x)$, yet merely confirms the bias by assigning low uncertainty to dark bands whose noise variance is genuinely low. Finally, SSIM [28], perceptual losses [15], and frequency-domain losses [17, 27, 29] operate uniformly over all pixels and are entirely blind to brightness-dependent noise structure.

Herein, the key difference of BBRD is that we explicitly normalize per-band error by empirical noise variance before reweighting, which is the critical step that transforms reweighting from a bias-amplifier into a bias-corrector.

Distributionally robust optimization. Group DRO [24] minimizes worst-group loss in classification and has strong convergence guarantees. However, applying Group DRO directly to image restoration requires non-trivial adaptation: unlike classification where groups are defined across samples, a single image patch spans multiple brightness levels, requiring groups to be defined at the pixel level within each patch. Furthermore, without noise normalization, naive Group DRO *worsens* brightness bias, as it mis-identifies bright bands as worst-performing because their raw MSE is inflated by signal-dependent noise, and upweights them further, compounding the exact problem we aim to fix.

In contrast, BBRD first normalizes per-band errors by empirical noise variance, ensuring the worst-performing band is identified on a noise-corrected scale rather than by raw loss magnitude, and then replaces the non-differentiable minimax objective with softmax DRO to enable dynamic reweighting during training. Without normalization, DRO upweights bright bands (highest raw loss); with it, DRO correctly upweights dark bands (highest noise-normalized loss).

3 Method

3.1 Preliminaries

Group Distributionally Robust Optimization. Standard ERM minimizes the average loss across all training samples, which can leave minority groups severely under-optimized. Group DRO [24] addresses this by minimizing the

worst-group loss instead:

$$\min_{\theta} \max_{g \in \mathcal{G}} \mathcal{L}_g(\theta),$$

where $\mathcal{G}$ is a predefined set of groups and $\mathcal{L}_g$ is the loss on group g. This ensures no group is left behind at the expense of aggregate performance.

3.2 Why MSE Fails: Two Compounding Factors

Although MSE is widely adopted, its uniform pixel treatment leads to a systematic gradient imbalance under signal-dependent noise. Two independent factors *compound* to over-emphasize bright pixels during training:

- **Signal-dependent residuals.** In early-to-mid training the residual approximates the noise, $r_i \approx n_i$, so $\mathbb{E}[r_i^2] \approx \mathrm{Var}(n_i)$ grows with brightness. Brighter pixels therefore produce larger gradient contributions via Eq. (1) by default, before any network-specific effects.
- **Brightness-correlated Jacobian.** The network Jacobian $\|\partial f_\theta(x_i)/\partial\theta\|$ is empirically larger for high-activation (bright) inputs across all 8 evaluated architectures—CNNs, Transformers, and State Space Models (Figs. 2a and 2b)—independently amplifying bright-pixel gradients.

Gradient analysis. The MSE gradient at pixel i is:

$$\left.\frac{\partial \mathcal{L}}{\partial \theta}\right|_i = \frac{2}{N}(f_\theta(x_i) - y_i) \cdot \frac{\partial f_\theta(x_i)}{\partial \theta}. \tag{1}$$

Proposition 1 (Gradient imbalance). *Under the following assumptions:* ***(A1)*** *residuals approximate noise in early-to-mid training ($r_i \approx n_i$);* ***(A2)*** *noise variance* $\mathrm{Var}(n_i)$ *and Jacobian norm* $\|\partial f_\theta(x_i)/\partial\theta\|$ *both increase with brightness (empirically confirmed across all 8 architectures in Figs. 2a and 2b), the expected squared gradient magnitude increases monotonically with brightness:*

$$\mathbb{E}\left[\left|\left.\frac{\partial \mathcal{L}}{\partial \theta}\right|_i\right|^2\right] \propto \underbrace{\mathbb{E}[r_i^2]}_{\text{factor (i)}} \cdot \underbrace{\mathbb{E}\left[\left\|\frac{\partial f_\theta(x_i)}{\partial \theta}\right\|^2\right]}_{\text{factor (ii)}}, \tag{2}$$

causing bright pixels to dominate parameter updates and leaving dark regions chronically under-optimized.

Proof. From Eq. (1), the squared gradient at pixel i is proportional to $r_i^2 \cdot \|J_i\|^2$. By (A1), $\mathbb{E}[r_i^2] \approx \mathrm{Var}(n_i)$ increases with brightness. By (A2), $\mathbb{E}[\|J_i\|^2]$ also increases with brightness. Since both non-negative factors increase monotonically with brightness, their product likewise increases monotonically, completing the argument. □

Jacobian measurement. $\|\partial f_\theta(x_i)/\partial\theta\|^2$ is approximated as $\sum_p \|\nabla_\theta^{(p)}\|_F^2$, the sum of squared Frobenius norms over all trainable parameter tensors p, computed via a single `backward()` pass on the per-pixel (or per-band) output scalar.

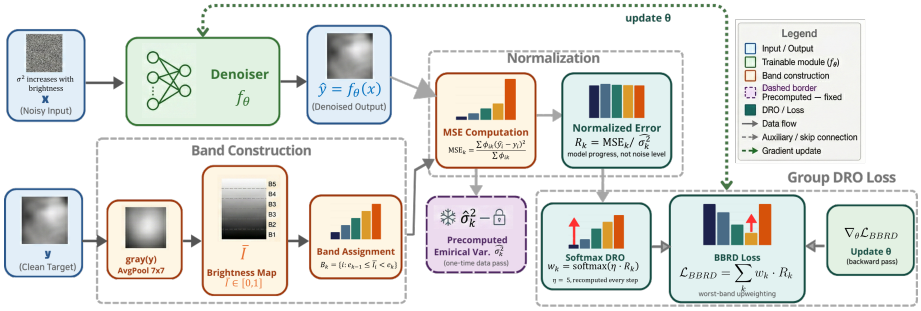


Fig. 3: BBRD training pipeline. A smoothed brightness map partitions pixels into K bands via GMM and Gaussian soft assignments (orange). Per-band MSE is normalized by empirical noise variance $\hat{\sigma}_k^2$, yielding R_k (teal). Softmax DRO (η=5) upweights the worst band, forming $\mathcal{L}_{\mathrm{BBRD}} = \sum_k w_k R_k$ (dark teal). Band computation is training-only; inference runs f_θ with zero overhead.

Per-band values are obtained by zeroing all output pixels outside band k before calling `backward()`, so that autograd accumulates gradients only from band-k pixels; the resulting $\sum_p \|\nabla_\theta^{(p)}\|_F^2$ is then divided by $|\mathcal{B}_k|$. No additional normalization is applied; Figs. 2a and 2b report these *unnormalized* per-band Jacobian norms as the primary result. As a sanity check, we also recompute norms normalized by parameter count per layer, $\tilde{J}_i = \sum_p \|\nabla_\theta^{(p)}\|_F^2 / |\theta^{(p)}|$, and confirm the same trend across all 8 architectures—including Transformer-based models (SwinIR, Restormer), whose self-attention layers could in principle redistribute activation energy, and State Space Models (MambaIR, MambaIRv2): the absolute scale changes but the brightest band consistently dominates the darkest in every case, confirming that the phenomenon is tied to brightness, not to layer size or architectural inductive bias. **Persistent bias.** This imbalance is self-reinforcing: once the network enters the fine-convergence regime ($r_i \ll n_i$), dark-pixel gradients are too weak to undo the accumulated deficit (Fig. 6), and any reweighting of raw MSE inherits this bias since the gradient signal is already corrupted by noise heteroscedasticity.

3.3 Brightness Bias-Robust Denoising (BBRD)

Overview: As illustrated in Fig. 3, BBRD corrects brightness bias via three sequential steps, each necessary: removing any one breaks the correction (Tab. 4). In the first step, we partition pixels into brightness-specific bands via a data-driven GMM with Gaussian soft assignments, enabling the loss to track brightness-specific failures smoothly across the full intensity range. In the second step, we normalize per-band error by empirical noise variance, placing all bands on a noise-corrected scale so that the truly worst-performing band, not merely the noisiest, is correctly identified. In the third step, we apply Softmax Group DRO to dynamically upweight whichever band is currently lagging, continu-

ously redirecting optimization pressure throughout training. The result is a zero-parameter, drop-in replacement for MSE with no inference overhead (Fig. 3).

Brightness Band Construction. Brightness bias manifests differently across the intensity spectrum: dark bands suffer from both low gradient magnitude and low noise variance, while bright bands dominate training due to the opposite. A single uniform MSE cannot track these band-specific failure modes; to redirect optimization pressure where it is truly needed, we must first partition pixels into brightness-specific groups so that per-band progress can be measured and compared on a common scale.

Brightness map. For a clean patch y, we compute a smoothed brightness map $\bar{I} = \mathrm{AvgPool}_{7\times7}(\mathrm{gray}(y)) \in [0,1]^{H\times W}$ using sRGB weights $(0.299, 0.587, 0.114)$. Bands are defined on the clean target y rather than the noisy input, avoiding noise-induced misassignments; this map is computed only during training, incurring no inference overhead.

GMM fitting and BIC selection. We fit a 1-D GMM to training-set intensities and select K^* via BIC, which balances fit quality against model complexity by penalizing the number of components. Unlike uniform binning, which arbitrarily splits the intensity range into equal intervals, GMM identifies where pixels naturally cluster in brightness space, placing band boundaries at the valleys between adjacent components where pixel density is lowest. This data-adaptive partitioning ensures each band captures a coherent group of pixels with similar brightness characteristics, rather than mixing pixels from different luminance modes. BIC then automatically selects the number of bands that best describes the data without overfitting, selecting K^*=5 for SIDD and K^*=3 for PolyU without manual tuning (Sec. 5.3). Then, we propose **Gaussian soft assignment.** Hard binning introduces gradient discontinuities at boundaries, where a small change in $\bar{I}_i$ abruptly switches band membership and can cause training instability. We instead assign pixel i to band k via a Gaussian soft weight:

$$\phi_{ik} = \exp\left(-\frac{(\bar{I}_i - c_k)^2}{2\sigma_g^2}\right), \quad c_k = \frac{e_{k-1} + e_k}{2}, \tag{3}$$

where e_{k-1} and e_k denote the valley boundaries flanking band k (i.e., the intensity minima between adjacent GMM components), c_k is the band center, and σ_g=0.05 ensures overlapping support near boundaries, so pixels near a boundary contribute smoothly to both adjacent bands. Note that ϕ_{ik} are unnormalized weights; normalization is implicit via the denominator $\sum_i \phi_{ik}$ in Eq. (4). Per-band MSE is then:

$$\mathrm{MSE}_k = \frac{\sum_i \phi_{ik}\,(\hat{y}_i - y_i)^2}{\sum_i \phi_{ik}}. \tag{4}$$

Empirical Noise Normalization. Without normalization, brighter bands always exhibit higher raw MSE simply due to signal-dependent noise, causing any reweighting scheme to misidentify them as the worst-performing and further amplify the bias we aim to correct. We normalize with per-band empirical noise variance $\hat{\sigma}_k^2$, precomputed in a single pass over the training set *before* training

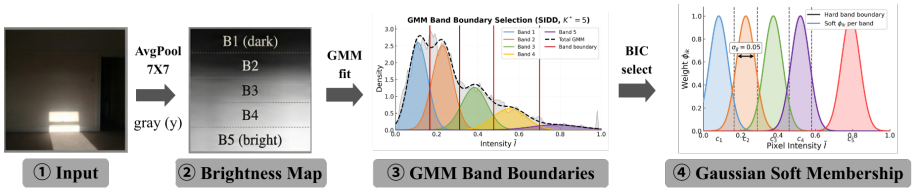


Fig. 4: Brightness band construction. (1) Input patch. **(2)** Smoothed brightness map $\bar{I}$ via AvgPool. **(3)** GMM fitted to training-set intensities; boundaries placed at inter-component valleys (K^*=5 for SIDD, K^*=3 for PolyU). **(4)** Gaussian soft weights ϕ_{ik} ensure smooth pixel-to-band assignment.

begins:

$$\hat{\sigma}_k^2 = \frac{\sum_i \phi_{ik}\,(x_i - y_i)^2}{\sum_i \phi_{ik}}, \qquad R_k = \frac{\mathrm{MSE}_k}{\hat{\sigma}_k^2}. \tag{5}$$

Interpretation: The normalized ratio R_k has a natural interpretation: $R_k \approx 1$ means no denoising progress; $R_k \to 0$ is perfect reconstruction. This requires no parametric noise model and adapts to any noise distribution. Since $\hat{\sigma}_k^2$ is a data property rather than a model parameter, it remains fixed throughout training.

Numerical stability: For very dark bands where $\hat{\sigma}_k^2$ is small, dividing by $\hat{\sigma}_k^2$ could amplify noise in R_k. We therefore clip $\hat{\sigma}_k^2$ from below by ϵ=10^{-6}:

$$R_k = \frac{\mathrm{MSE}_k}{\max(\hat{\sigma}_k^2,\, \epsilon)}.$$

In practice, $\hat{\sigma}_k^2$ never falls below 10^{-4} on either dataset (B1: $2.2{\times}10^{-3}$ on SIDD), so this safeguard is never triggered but ensures robustness in out-of-distribution settings.

Softmax Group DRO. After noise normalization, bands still improve at different rates throughout training: bright bands converge quickly while dark bands lag persistently, and the gap between them changes at every step. A static weight preset before training cannot respond to this ongoing disparity; only a mechanism that recomputes weights at every training step can continuously redirect gradient flow to whichever band is currently most in need.

We approximate the non-differentiable minimax $\min_\theta \max_k R_k$ with differentiable softmax reweighting over valid bands $\mathcal{V}$=$\{k : \sum_i \phi_{ik} > 0\}$:

$$\mathcal{L}_{\mathrm{BBRD}} = \sum_{k\in\mathcal{V}} w_k \cdot R_k, \quad w_k = \frac{\exp(\eta R_k)}{\sum_{j\in\mathcal{V}} \exp(\eta R_j)}, \tag{6}$$

where η=5 uniformly across all backbones and datasets.

Temperature η: As $\eta \to \infty$, the loss approaches the hard minimax $\max_k R_k$; at η=0, it reduces to a uniform average of R_k (normalized MSE). The value η=5 achieves the balance between worst-band focus and overall performance, and remains stable across $\eta \in [0.5, 10]$ (Sec. 5). Crucially, w_k are recomputed at every training step: as a lagging band improves, its weight decreases and gradient flow shifts to the next worst band (Fig. 6).

3.4 Per-Band Evaluation Protocol

PSNR-D and **PSNR-B** denote per-band PSNR on the darkest and brightest regions, respectively. For SIDD, we use $[0, 0.2)$ / $[0.8, 1.0]$; for PolyU, whose brightness is heavily skewed toward bright values (bright/dark ratio $\approx$18.5$\times$), we use $[0, 0.45)$ / $[0.7, 1.0]$ to ensure sufficient dark-band pixel counts. BBRD consistently outperforms MSE across all alternative dark-band thresholds: on SIDD, gains range from +0.45 dB at $[0, 0.3)$ to +0.58 dB at $[0, 0.15)$, with the largest gain at the most extreme dark region, confirming that BBRD is most effective precisely where brightness bias is most severe (full threshold sweep in supplementary).

4 Experiments

Datasets. We evaluate on two complementary real-world benchmarks. SIDD [1] consists of smartphone images skewed toward dark values (Fig. 1a), making it a natural testbed for dark-band failures. PolyU [31] provides DSLR images with an extreme bright-to-dark energy ratio of 18.5$\times$, a complementary regime where dark pixels are sparse and easily overlooked under aggregate evaluation.
Architectures. To verify that brightness bias is a loss-level phenomenon rather than a model-specific artifact, we train 8 architectures spanning CNNs (NAFNet [4], DRUNet [34], KBNet [36], SCUNet [33]), Transformers (SwinIR [19], Restormer [32]), and State Space Models (MambaIR [10], MambaIRv2 [9]).
Training. All models are trained with identical settings (Adam, 100 epochs, fp32; full hyperparameters in supplementary). BBRD requires *no* per-backbone tuning, as K, η, and σ_g are fixed across all 8 architectures and both datasets. It introduces only +1.7% training overhead, broken down as follows: GMM fitting and $\hat{\sigma}_k^2$ estimation are performed *once* before training ($\sim$2 min on SIDD); per-step, the soft weights ϕ_{ik} are recomputed from the clean target's brightness map (element-wise Gaussian, negligible), and the softmax DRO adds K scalar exponentials and one normalization. BBRD incurs **zero inference cost**: band assignments and DRO weights are discarded after training, leaving the deployed model bitwise identical to one trained with vanilla MSE. BBRD operates in sRGB to match both datasets and all evaluated architectures, with $\hat{\sigma}_k^2$ estimated empirically from training pairs (x, y), automatically absorbing ISP nonlinearities without any parametric noise assumption. $\hat{\sigma}_k^2$ is stable across subsamples (CV <15%), ISO subgroups (<8%), and cross-dataset transfer (SIDD $\hat{\sigma}_k^2$ on PolyU yields PSNR-D within 0.1 dB). A 2$\times$ uniform scaling of $\hat{\sigma}_k^2$ changes PSNR-D by <0.3 dB, consistent with the observation that DRO weights depend on relative band ranking rather than absolute $\hat{\sigma}_k^2$ values (supplementary).
Baselines. We compare against 13 loss functions across four paradigms: pixel-wise regression (ℓ_2, ℓ_1, Charbonnier [3], Huber [12]), structural similarity (SSIM [28], MS-SSIM [29]), frequency-domain objectives (FFT [27], ℓ_1+FFT [8], Focal Freq. [14], Guided Freq. [2]), and sample reweighting (Focal [20], OHEM [25], Heteroscedastic NLL [16]).

Table 1: Loss comparison (NAFNet). PSNR-D / PSNR-B: SIDD $[0, 0.2)$ / $[0.8, 1.0]$; PolyU $[0, 0.45)$ / $[0.7, 1.0]$. Every baseline either preserves or worsens the dark/bright disparity; BBRD is the only method that improves both bands simultaneously. Best **bold**, second-best underlined.

	SIDD (smartphone)				**PolyU** (DSLR)			
Method	PSNR-D↑	PSNR-B↑	PSNR↑	SSIM↑	PSNR-D↑	PSNR-B↑	PSNR↑	SSIM↑
Pixel-wise regression								
MSE (ℓ_2)	37.51	38.37	39.47	.9389	35.80	37.82	38.50	.9748
ℓ_1	37.60	37.67	39.65	.9390	35.94	37.87	38.68	.9783
Charbonnier [3]	37.48	38.66	39.65	.9398	35.99	38.18	38.79	.9794
Huber [12]	37.44	38.34	39.51	.9400	35.90	37.65	38.35	.9754
Structural								
SSIM [28]	36.81	36.49	38.78	.9337	36.07	38.05	38.85	.9793
MS-SSIM [29]	37.50	37.64	39.36	.9396	36.06	37.90	38.72	.9779
Frequency-domain								
FFT [27]	37.64	37.96	39.59	.9406	35.84	37.56	38.44	.9764
ℓ_1+FFT [8]	37.70	38.28	39.80	.9417	36.11	38.20	38.80	.9786
Focal Freq. [14]	36.10	37.07	37.67	.9058	35.20	37.15	37.54	.9622
Guided Freq. [2]	37.57	38.60	39.50	.9389	36.15	38.14	38.92	.9802
Reweighting & mining								
Focal [20]	37.19	38.55	39.16	.9339	35.54	37.48	38.07	.9711
OHEM [25]	37.47	38.64	39.44	.9385	36.08	37.92	38.62	.9779
Heteroscedastic NLL [16]	37.29	38.09	39.37	.9386	36.09	38.12	38.75	.9788
Ours								
BBRD	**37.96**	**38.69**	**40.12**	**.9442**	**36.27**	**38.34**	**39.06**	**.9804**

4.1 Main Results

Tab. 1 reveals a consistent failure: every baseline either preserves or worsens the dark/bright disparity, as raw residuals inflated by signal-dependent noise cause all pixel-wise objectives to systematically over-serve bright bands. Notably, the closest prior work—Heteroscedastic NLL [16]—*worsens* PSNR-D below MSE (37.29 vs. 37.51 dB), as its learned per-pixel $\sigma(x)$ assigns low uncertainty to dark bands whose noise variance is genuinely low, reinforcing the bias. BBRD addresses this by normalizing at the *group level*, a distinction per-pixel methods cannot capture.

BBRD breaks this trade-off: in our experiments it is the **only method among all 13 alternatives that simultaneously improves PSNR-D and PSNR-B on both datasets**: **+0.65 dB** aggregate PSNR on SIDD and **+0.56 dB** on PolyU, with per-band gains reaching up to **+0.89 dB**, pushing the PSNR-D vs. PSNR-B Pareto frontier strictly outward (Fig. 2c).

Tab. 2 confirms this is a loss-level bottleneck: BBRD improves every metric on all 8 backbones—CNN, Transformer, and SSM—with no per-backbone tuning. Qualitative comparisons (Fig. 5) show BBRD corrects dark-region residuals that MSE leaves behind.

Table 2: Architecture generalization (MSE vs. BBRD). BBRD improves all metrics on all 8 backbones with no per-backbone tuning. PSNR-D / PSNR-B: SIDD $[0, 0.2)$ / $[0.8, 1.0]$; PolyU $[0, 0.45)$ / $[0.7, 1.0]$.

		SIDD				PolyU			
Backbone	Loss	PSNR-D↑	PSNR-B↑	PSNR↑	SSIM↑	PSNR-D↑	PSNR-B↑	PSNR↑	SSIM↑
NAFNet	MSE	37.51	38.37	39.47	.9389	35.80	37.82	38.50	.9748
	BBRD	**37.96**	**38.69**	**40.12**	**.9442**	**36.27**	**38.34**	**39.06**	**.9804**
DRUNet	MSE	38.02	38.63	40.04	.9445	35.92	37.98	38.50	.9755
	BBRD	**38.35**	**39.31**	**40.46**	**.9473**	**36.40**	**38.12**	**39.07**	**.9800**
KBNet	MSE	37.62	38.48	39.82	.9412	36.28	38.34	38.89	.9749
	BBRD	**38.43**	**38.76**	**40.41**	**.9462**	**36.47**	**38.77**	**39.00**	**.9793**
SwinIR	MSE	37.29	39.27	39.48	.9422	36.28	38.03	39.01	.9792
	BBRD	**37.83**	**39.77**	**40.23**	**.9450**	**36.33**	**38.29**	**39.23**	**.9798**
Restormer	MSE	37.88	38.66	40.19	.9437	36.33	38.54	39.01	.9782
	BBRD	**38.56**	**39.00**	**40.72**	**.9485**	**36.64**	**38.67**	**39.10**	**.9800**
SCUNet	MSE	38.26	38.82	40.34	.9461	35.96	38.08	38.68	.9772
	BBRD	**38.42**	**39.37**	**40.54**	**.9467**	**36.24**	**38.24**	**39.13**	**.9796**
MambaIR	MSE	37.96	38.93	40.09	.9432	36.16	38.09	38.77	.9773
	BBRD	**38.22**	**39.82**	**40.26**	**.9467**	**36.32**	**38.30**	**39.10**	**.9791**
MambaIRv2	MSE	38.14	39.35	40.04	.9449	36.02	38.11	38.75	.9758
	BBRD	**38.45**	**39.72**	**40.52**	**.9476**	**36.32**	**38.29**	**39.22**	**.9791**

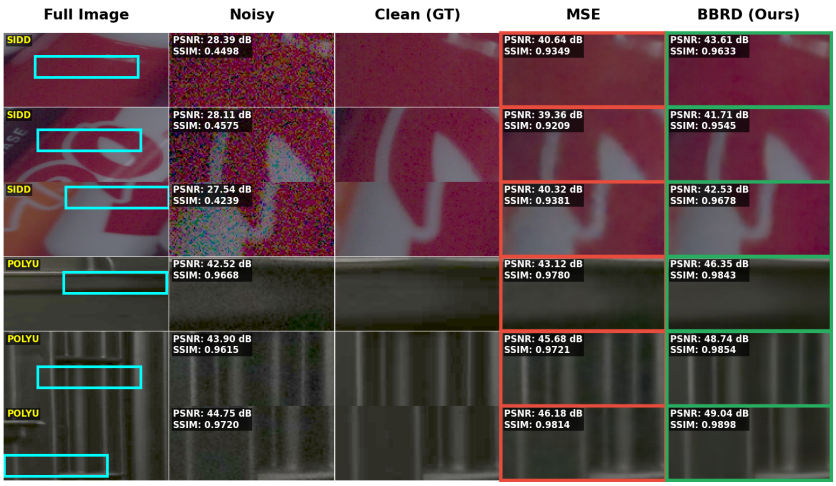


Fig. 5: Qualitative comparison (SIDD and PolyU). BBRD corrects dark-region residuals left behind by MSE (green = improvement; red = regression, confined to bright areas where MSE already performs well).

Table 3: LumaSet-700 results (SIDD-trained NAFNet, averaged over all 11 levels). BBRD improves PSNR-D, PSNR-B, PSNR, and SSIM across all three noise types. PSNR-D / PSNR-B: $[0, 0.2)$ / $[0.8, 1.0]$. Per-level curves are in the supplementary.

		Sig.-dep.				Random				Gaussian			
	Loss	PSNR-D↑	PSNR-B↑	PSNR↑	SSIM↑	PSNR-D↑	PSNR-B↑	PSNR↑	SSIM↑	PSNR-D↑	PSNR-B↑	PSNR↑	SSIM↑
NAFNet	MSE	26.50	19.00	26.43	.6536	26.52	21.37	27.29	.6887	26.68	21.50	27.44	.6931
	BBRD	**26.80**	**19.37**	**26.94**	**.6870**	**26.81**	**21.76**	**27.71**	**.7123**	**26.98**	**21.89**	**27.87**	**.7172**

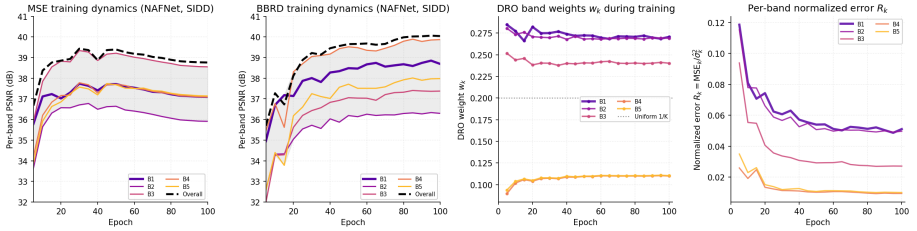


Fig. 6: Training dynamics and DRO stability. (a) MSE locks in the dark/bright gap early. **(b)** BBRD closes this gap across all bands. **(c)** DRO weights stabilize with B1 (dark) highest. **(d)** R_k decreases monotonically on a noise-corrected scale.

5 Ablation Studies

5.1 Noise Type Robustness

We evaluate the SIDD-trained NAFNet on **LumaSet-700**, a synthetic dataset of 700 patches stratified uniformly across five brightness quintiles ($[0, 0.2)$, $[0.2, 0.4)$, $\ldots$, $[0.8, 1.0]$), each containing 140 patches cropped from publicly available natural images and rendered under three noise types (signal-dependent SIDD-style, homoscedastic Gaussian, and random-variance) at 11 severity levels (L0 mild $\rightarrow$ L10 max). This is the only benchmark with uniform $[0, 1]$ brightness coverage and controllable noise. LumaSet-700 will be released upon publication. BBRD outperforms MSE across all noise types and levels (Tab. 3): the largest gain appears under signal-dependent noise ($+0.51$ dB), where variance normalization is most principled, while gains persist even under homoscedastic noise—confirming that BBRD addresses both compounding sources of brightness bias, not just noise heteroscedasticity. Full per-level curves are in the supplementary.

5.2 Training Dynamics and DRO Stability

Fig. 6 traces the full training trajectory. Under MSE (a), the dark/bright disparity locks in within the first few epochs and never recovers. BBRD (b) closes this gap from epoch 1, surpassing MSE's worst-band ceiling on *every* band by epoch 100 (gains: $+$**0.55 dB** darkest $\rightarrow$ $+$**1.02 dB** mid-tones). Band weights w_k (c) converge and stabilize with no oscillation; normalized error R_k (d) decreases monotonically, confirming that DRO continuously tracks the truly lagging band on a noise-corrected scale throughout training.

Table 4: Component ablation (NAFNet, SIDD). DRO without normalization worsens PSNR-D below MSE; normalization must precede DRO. η=0 (normalization only) also fails, confirming both components are jointly necessary. Default in gray; ↓ denotes below MSE. PSNR-D / PSNR-B: $[0, 0.2)$ / $[0.8, 1.0]$.

Component	Setting	PSNR↑	SSIM↑	PSNR-D↑	PSNR-B↑
Build-up	MSE baseline	39.47	.9389	37.51	38.37
	DRO only (no norm, η=5)	39.43	.9402	36.67↓	38.44
	+ GMM bands, no $\hat{\sigma}_k^2$	39.81	.9420	37.07↓	38.50
	Full BBRD (GMM+Emp+Gaussian)	**40.12**	**.9442**	**37.96**	**38.69**
Simple baselines	Per-pixel inv. variance weighting	39.11	.9382	37.35↓	38.21↓
	Uniform quantile bands + norm + DRO	39.97	.9426	37.91	38.70
	GMM bands + norm + DRO (ours)	**40.12**	**.9442**	**37.96**	**38.69**
Band source	Band from noisy x	39.03	.9381	37.43↓	38.40↓
	Band from clean y (ours)	**40.12**	**.9442**	**37.96**	**38.69**
DRO temperature η	η=0 *(norm only, no DRO)*	39.53	.9373	36.82↓	38.51
	η=1	40.04	.9437	36.80↓	38.67
	η=5 (ours)	**40.12**	**.9442**	**37.96**	**38.69**

5.3 Component Ablation

Tab. 4 traces a bottom-up path from MSE to full BBRD. Crucially, normalization must precede DRO: without it, DRO *worsens* PSNR-D below MSE (36.67 dB↓) by misidentifying high-noise bright bands as worst-performing. Similarly, η=0 (normalization only) fails (36.82 dB↓), confirming both components are jointly necessary. Band assignments from noisy x rather than clean y drop PSNR-D by 0.53 dB, and uniform quantile bands fall 0.05 dB short, validating our GMM-based clean-target design. K=5 (BIC-selected) outperforms K=4 (-0.24 dB) and K=8 (-0.30 dB); full sensitivity analysis is in the supplementary.

6 Conclusion

We propose Brightness Bias-Robust Denoising (BBRD), a plug-and-play module that tackles an overlooked problem in image denoising tasks: dark pixels are reconstructed substantially worse than the bright ones despite carrying less signal-dependent noise. Specifically, BBRD contains three stages. First, GMM band partitioning with Gaussian soft assignments identifies brightness-specific failure modes in a data-adaptive manner. Second, empirical noise normalization places all bands on a noise-corrected scale, ensuring the truly underperforming band (not merely the noisiest) is correctly identified. Finally, softmax Group DRO dynamically redirects gradient flow to whichever band is currently lagging throughout training, with zero additional parameters or inference cost. Experiments across 8 architectures and both datasets demonstrate that, in our experiments, BBRD is the only objective among 13 alternatives that improves every brightness band simultaneously, delivering up to $+$**0.45 dB** on dark bands, $+$**0.32 dB** on bright bands, and $+$**0.65 dB** aggregate PSNR on SIDD (NAFNet).

Acknowledgements

This work was supported by research grants from the RGC (Project Nos. 16500825, R6005-24, and C6088-25Y), a research grant from the Joint Research Scheme (JRS) under the National Natural Science Foundation of China (NSFC) and the Research Grants Council (RGC) of Hong Kong (Project No. N_HKUST654/24), and a research grant from the Innovation and Technology Commission (ITC) (Project No. GHP/124/22).

References

1. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: CVPR (2018) 10
2. Benjdira, B., Ali, A.M., Koubaa, A.: Guided frequency loss for image restoration. arXiv preprint arXiv:2309.15563 (2023) 10, 11
3. Charbonnier, P., Blanc-Féraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: ICIP. vol. 2, pp. 168–172 (1994) 5, 10, 11
4. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: ECCV (2022) 1, 2, 4, 10
5. Chen, Z., Zhang, J., Lai, Z., Zhu, G., Liu, Z., Chen, J., Li, J.: The devil is in the crack orientation: A new perspective for crack detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision 1
6. Dong, L., Fan, Q., Guo, Y., Wang, Z., Zhang, Q., Chen, J., Luo, Y., Zou, C.: TSD-SR: One-step diffusion with target score distillation for real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23174–23184 (Jun 2025) 4
7. Elad, M., Kawar, B., Vaksman, G.: Image denoising: The deep learning revolution and beyond—A survey paper. SIAM Journal on Imaging Sciences **16**(3), 1594–1654 (2023). https://doi.org/10.1137/23M1545859 1
8. Fu, Y., et al.: NTIRE 2024 challenge on low light image enhancement: Methods and results. In: CVPRW (2024) 10, 11
9. Guo, H., Guo, Y., Zha, Y., Zhang, Y., Li, W., Dai, T., Xia, S.T., Li, Y.: MambaIRv2: Attentive state space restoration. In: CVPR (2025) 1, 4, 10
10. Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.T.: MambaIR: A simple baseline for image restoration with state-space model. In: ECCV (2024) 1, 4, 10
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020), arXiv:2006.11239 4
12. Huber, P.J.: Robust estimation of a location parameter. The Annals of Mathematical Statistics **35**(1), 73–101 (1964) 5, 10, 11
13. Jiang, B., et al.: Eficient image denoising using deep learning: A brief survey. Information Fusion (2025). https://doi.org/10.1016/j.inffus.2025.100867 1
14. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: ICCV (2021) 10, 11
15. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016) 5
16. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: NeurIPS (2017) 5, 10, 11

17. Khor, H.Q., Jiang, J., Ong, E.J.: Wavelet-based dual-branch network for image demoireing. In: ECCV (2022) 5
18. Li, F., Li, J., Conde, M., et al.: AIM 2025 challenge on real-world RAW image denoising. In: ICCVW (2025), arXiv:2510.06601 1
19. Liang, J., Cao, J., Sun, G., Zhang, K., Gool, L.V., Timofte, R.: SwinIR: Image restoration using swin transformer. In: ICCVW (2021) 1, 4, 5, 10
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017) 3, 5, 10, 11
21. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: DiffBIR: Toward blind image restoration with generative diffusion prior. In: ECCV (2024), arXiv:2308.15070 4
22. Mao, J., Sun, L., Chen, J., Yu, S.: Overview of research on digital image denoising methods. Sensors **25**(8), 2615 (Apr 2025). https://doi.org/10.3390/s25082615, pMCID: PMC12031399 2
23. Pathak, K., Bhandari, A.K.: Advances in deep learning and filtering models for medical image denoising: A review of current and future trends. Circuits, Systems, and Signal Processing (2026). https://doi.org/10.1007/s00034-025-03477-z 1
24. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In: ICLR (2020) 5
25. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: CVPR (2016) 3, 5, 10, 11
26. Sun, L., Guo, H., et al.: The tenth NTIRE 2025 image denoising challenge report. In: CVPRW (2025), arXiv:2504.12276 1
27. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general U-shaped transformer for image restoration. In: CVPR (2022) 5, 10, 11
28. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004) 5, 10, 11
29. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Asilomar Conference on Signals, Systems and Computers. pp. 1398–1402 (2003) 5, 10, 11
30. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. arXiv preprint arXiv:2406.08177 (2024). https://doi.org/10.48550/arXiv.2406.08177 4
31. Xu, J., Li, H., Liang, Z., Zhang, D., Zhang, L.: Real-world noisy image denoising: A new benchmark. arXiv preprint arXiv:1804.02603 (2018) 10
32. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR (2022) 1, 4, 5, 10
33. Zhang, K., Li, Y., Liang, J., Cao, J., Zhang, Y., Tang, H., Fan, D.P., Timofte, R., Gool, L.V.: Practical blind image denoising via Swin-Conv-UNet and data synthesis. Machine Intelligence Research (2023). https://doi.org/10.1007/s11633-023-1466-0 10
34. Zhang, K., Li, Y., Zuo, W., Zhang, L., Gool, L.V., Timofte, R.: Plug-and-play image restoration with deep denoiser prior. IEEE TPAMI (2021). https://doi.org/10.1109/TPAMI.2021.3088914 1, 4, 10
35. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing **26**(7), 3142–3155 (2017). https://doi.org/10.1109/TIP.2017.2662206 1, 4

36. Zhang, Y., Li, D., Shi, X., He, D., Song, K., Wang, X., Qin, H., Li, H.: KBNet: Kernel basis network for image restoration. arXiv preprint arXiv:2303.02881 (2023) 10