Effect of Noise Level and Brightness/Contrast on Image Denoising Performance: Classical vs. Neural Network Denoisers

Andrea Faccioli
Mid Sweden University
Sundsvall, Sweden
anfa2502@student.miun.se

***Abstract*—Reliable image restoration is critical for modern visual data analysis, yet the performance of denoising algorithms under varying photometric conditions remains under-explored. This study systematically investigates the combined effects of noise intensity, brightness, and contrast on the performance of classical and neural network-based denoisers. Using the Kodak24 dataset, BM3D and the deep learning-based DRUNet have been evaluated across a procedurally manipulated set of images degraded by both Additive White Gaussian Noise and Poisson noise. Performance is assessed using pixel-wise fidelity (PSNR) and deep perceptual similarity (LPIPS). Statistical analysis via ANOVA (Analysis of Variance) and Linear Mixed Models reveals that while the neural denoiser exhibits superior perceptual stability at high noise levels, a significant divergence exists between metrics. Increased brightness and contrast were found to improve perceptual quality while paradoxically degrading signal fidelity scores, underscoring the inability of standard metrics to accurately penalize texture loss in low-light environments. Furthermore, interaction effects demonstrate that high brightness mitigates Gaussian noise effectively but offers diminishing returns for signal-dependent Poisson noise. These findings underscore the necessity of multi-metric evaluation for developing adaptive restoration pipelines.**

***Index Terms*—image denoising, bm3d, drunet, lpips, contrast, brightness**

I. Introduction

The rapid proliferation of digital imaging devices, ranging from high-end smartphones and medical sensors to resource-constrained Internet-of-Things (IoT) cameras and surveillance systems, has established visual data as a cornerstone of modern analysis and decision-making. As the volume of image capture increases, the reliability of the underlying image processing pipelines becomes critical for practical applications.

Despite significant advancements in sensor technology, image degradation remains a pervasive challenge. While premium hardware often integrates sophisticated software pipelines to minimize artifacts, a vast number of low-cost or embedded sensors continue to produce images compromised by significant noise. This issue is frequently exacerbated by environmental factors, particularly under challenging illumination conditions where the interplay between noise, low brightness, and poor contrast severely impacts perceived image quality [1].

Over the last several decades, researchers have proposed hundreds of methods to tackle this problem. However, while the general efficacy of these methods is well-documented, their comparative robustness when subjected to simultaneous variations in noise intensity and photometric properties (brightness and contrast) requires further investigation.

Understanding these interactions is vital for developing adaptive algorithms capable of performing reliably across diverse real-world imaging conditions. Consequently, this study systematically investigates how noise level, brightness, and contrast influence the performance of both classical and neural network-based image denoisers. By analyzing the denoising results across a manipulated dataset, this research aims to provide answers to the following questions:

- **RQ1**: How does the rate of performance degradation of neural network–based denoisers compare with classical methods as noise levels increase?
- **RQ2**: To what extent do brightness/contrast changes improve or degrade image denoising performance?
- **RQ3**: How do brightness/contrast changes interact with noise type to influence denoiser performance?

II. Related work

Image denoising has evolved from local spatial filters to non-local transform methods and data-driven neural networks [2].

Classical Denoising Approaches: Early local smoothing methods often blurred edges, prompting the development of Non-Local Means (NLM) [3], which exploited recurring global patterns. Building on this, Block-Matching and 3D Filtering (BM3D) [4] groups similar patches for collaborative transform-domain filtering. While BM3D remains a benchmark for structural preservation via sparsity constraints, it is computationally intensive and limited by fixed, handcrafted priors.

Neural Network-Based Denoising: Deep learning introduced effective learning-based solutions. Early CNNs like DnCNN [5] used residual learning to subtract noise, outperforming classical techniques. More recently, the Deep Residual U-Net (DRUNet) [6] combined U-Net multi-scale features with residual connections for plug-and-play restoration. Unlike BM3D, DRUNet learns complex priors from vast datasets, offering superior generalization across varying noise conditions.

III. Experimental setup

A. Dataset and pre-processing

The experiments were conducted using the Kodak24 image dataset, a widely adopted benchmark for image restoration and denoising research. The dataset consists of 24 natural RGB images with a resolution of 768×512 pixels, covering a diverse range of scenes, textures, and illumination conditions. These characteristics make Kodak24 suitable for analyzing how denoisers behave across different visual structures.

To investigate the combined effect of brightness, contrast, and noise on denoising performance, each image was systematically manipulated along three dimensions:

- **Brightness levels:** low, original, high
- **Contrast levels:** low, original, high
- **Noise type and level:**
  - Additive White Gaussian Noise (AWGN) at three intensities (low, medium, high)
  - Poisson noise at three intensities (low, medium, high)

Specifically, the brightness and contrast adjustments were implemented using the `ImageEnhance` module from the Python PIL (Pillow) library. A mapping between categorical variables (i.e. low, original, high) and the actual parameters values can be found in Table I

TABLE I: Summary of Image Manipulation Parameters

Dimension	Level	Parameter Value
Brightness	Low	Factor 0.5
	Original	Factor 1.0
	High	Factor 2.0
Contrast	Low	Factor 0.5
	Original	Factor 1.0
	High	Factor 2.0
Noise (AWGN)	Low	$\sigma = 10$
	Medium	$\sigma = 30$
	High	$\sigma = 70$
Noise (Poisson λ = Photon Scale)	Low Noise	Peak $\lambda = 100$
	Medium Noise	Peak $\lambda = 30$
	High Noise	Peak $\lambda = 10$

This procedure yields nine brightness–contrast combinations for each original image. Each of these combinations was then corrupted with six noise configurations (two types × three levels), resulting in 54 manipulated versions per image.

Across the full dataset, this produced a total of 1,296 corrupted images, each of which was subsequently processed by two selected denoisers. A visualization of the complete image manipulation pipeline can be found in Fig. 1. In total, 2,592 denoised outputs were generated and used for the quantitative analysis described in Section III-D.

B. Selected denoisers

To ensure a representative comparison between classical and modern image denoising paradigms, two complementary denoisers were selected for this study: **BM3D** and **DRUNet**.

However, it is important to note that both BM3D and DRUNet are primarily designed to handle Additive White Gaussian Noise (AWGN). To accommodate the Poisson noise configurations used in this study, the Anscombe transform [7] was applied to the Poisson-degraded images prior to denoising.

Together, BM3D and DRUNet allow for a balanced evaluation of classical versus deep learning approaches, enabling an in-depth analysis of how different algorithmic families respond to variations in noise, brightness, and contrast.

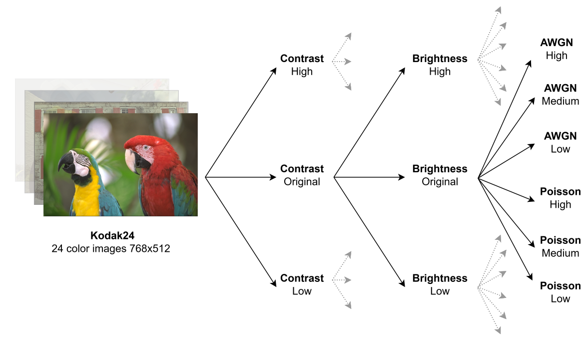

Fig. 1: Visualization of the image manipulation pipeline

C. Selected performance metrics

To ensure a balanced evaluation of both signal fidelity and perceptual quality, two complementary metrics were employed: Peak Signal-to-Noise Ratio (PSNR) and Learned Perceptual Image Patch Similarity (LPIPS) [8].

PSNR provides a standard quantitative measure of pixel-wise reconstruction error. However, as pixel-fidelity often diverges from human visual assessment, **LPIPS** is utilized to measure perceptual distance. By comparing high-level features extracted from pre-trained deep neural networks, LPIPS captures structural and textural similarities that align more closely with human perception than traditional Euclidean metrics.

D. Statistical analysis

Statistical computations were performed using the R environment.

To address **RQ1**, a two-way Analysis of Variance for Repeated Measures (RM-ANOVA) was conducted to evaluate the main effects and interaction between `Denoiser Type` and `Noise Level`, specifically testing if degradation rates differ significantly between classical and neural methods.

To address **RQ2** and **RQ3**, Linear Mixed Models (LMM) were employed to account for the repeated measures design (multiple manipulations applied to the same 24 source images). The LMM specification included:

- Fixed Effects: `Denoiser Type`, `Noise Type`, `Noise Level`, `Brightness`, `Contrast` and a subset of their interactions.
- Random Effects: A random intercept for `image_idx` to control for the intrinsic variability and difficulty of specific source images.

IV. Results

The experimental results on the degradation rate of performance (RQ1) are visualized in Fig. 2. While PSNR degradation appears similar for both denoisers, the LPIPS trajectories diverge. This observation is statistically confirmed by the ANOVA tests in Table II, where the interaction term `Denoiser type` × `Noise level` is significant for LPIPS ($F = 19.25$, $p < 0.001$), but not for PSNR.

Table III details the Linear Mixed Model analysis regarding impact of brightness and contrast (RQ2). The (Intercept) establishes the baseline performance for BM3D under low

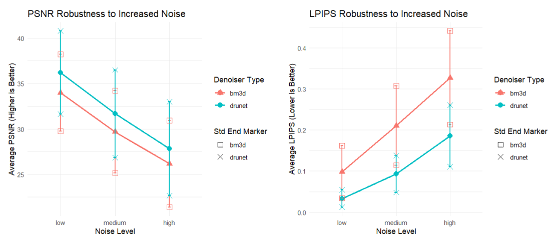

Fig. 2: PSNR and LPIPS plots. X axis is the noise level, Y axis is the metric value. Error bars denote ± 1 standard deviation from the mean

TABLE II: Repeated Measures ANOVA results for PSNR and LPIPS.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Source (PSNR)	Df	Sum Sq	Mean Sq	F
Denoiser type	1	745	745	33.73***
Noise level	2	6640	3320	150.19***
Denoiser type × noise level	2	7	3.5	0.15
Residuals	2580	57028	22	

Source (LPIPS)	Df	Sum Sq	Mean Sq	F
Denoiser type	1	2.010	2.010	350.96***
Noise level	2	3.681	1.841	321.36***
Denoiser type × noise level	2	0.221	0.110	19.25***
Residuals	2580	14.776	0.006	

TABLE III: LMM results for PSNR and LPIPS

Fixed effect (PSNR)	Estimate	Std. Error	t value
(Intercept)	42.309	0.344	122.86
Denoiser: DRUNet	1.961	0.074	26.64
Noise type: Poisson	-2.931	0.074	-39.81
Brightness: original	-3.516	0.090	-38.98
Brightness: high	-6.904	0.090	-76.56
Contrast: original	-3.854	0.090	-42.74
Contrast: high	-5.899	0.090	-65.41
Noise level: medium	-4.426	0.090	-49.08
Noise level: high	-8.110	0.090	-89.93
Fixed effect (LPIPS)	**Estimate**	**Std. Error**	**t value**
(Intercept)	0.177	0.008	22.51
Denoiser: DRUNet	-0.108	0.002	-45.24
Noise type: Poisson	-0.005	0.002	-2.25
Brightness: original	-0.022	0.003	-7.49
Brightness: high	-0.051	0.003	-17.32
Contrast: original	-0.026	0.003	-9.04
Contrast: high	-0.066	0.003	-22.48
Noise level: medium	0.086	0.003	29.43
Noise level: high	0.191	0.003	65.24

AWGN noise, low brightness, and low contrast. For both brightness and contrast, the 'original' and 'high' levels yield highly significant negative coefficients ($|t| > 7$). Given the metric definitions (higher PSNR is better; lower LPIPS is better), these negative estimates indicate divergent outcomes: increased brightness and contrast degrade PSNR scores (e.g., -6.90 for High Brightness) while simultaneously improving perceptual LPIPS scores (e.g., -0.051 for High Brightness). The underlying causes of this contrasting behavior between signal fidelity (PSNR) and perceptual quality (LPIPS) are further analyzed in the Discussion section.

Regarding the interaction between photometric changes and noise type (RQ3), Table IV highlights another disparity. The interaction term `Brightness: high × Noise type: Poisson` reveals the most severe degradation in the entire

TABLE IV: LMM results PSNR and LPIPS with interactions

Fixed effect (PSNR)	Estimate	Std. Error	t value
(Intercept)	41.397	0.345	119.86
Denoiser: DRUNet	1.961	0.062	31.64
Noise level: medium	-4.426	0.076	-58.30
Noise level: high	-8.110	0.076	-106.83
Brightness: orig	-3.046	0.107	-28.37
Brightness: high	-4.577	0.107	-42.63
Contrast: orig	-3.973	0.107	-37.01
Contrast: high	-5.839	0.107	-54.39
Noise type: Poisson	-1.106	0.139	-7.98
Brightness: orig × Noise type: Poisson	-0.940	0.152	-6.19
Brightness: high × Noise type: Poisson	-4.654	0.152	-30.65
Contrast: orig × Noise type: Poisson	0.238	0.152	1.57
Contrast: high × Noise type: Poisson	-0.119	0.152	-0.79
Fixed effect (LPIPS)	**Estimate**	**Std. Error**	**t value**
(Intercept)	0.179	0.008	21.81
Denoiser: DRUNet	-0.108	0.002	-45.68
Noise level: medium	0.086	0.003	29.72
Noise level: high	0.191	0.003	65.87
Brightness: orig	-0.027	0.004	-6.70
Brightness: high	-0.068	0.004	-16.65
Contrast: orig	-0.020	0.004	-4.87
Contrast: high	-0.054	0.004	-13.25
Noise type: Poisson	-0.009	0.005	-1.66
Brightness: orig × Noise type: Poisson	0.011	0.006	1.91
Brightness: high × Noise type: Poisson	0.035	0.006	6.06
Contrast: orig × Noise type: Poisson	-0.013	0.006	-2.24
Contrast: high × Noise type: Poisson	-0.023	0.006	-3.96

dataset for signal fidelity, with a PSNR estimate of -4.654 ($t = -30.65$). This specific interaction also penalizes the perceptual quality (LPIPS estimate $+0.035$, $t = 6.06$).

In contrast, the interaction `Contrast: high × Noise type: Poisson` shows no statistically significant impact on pixel fidelity (PSNR $t = -0.79$). However, this same interaction yields a statistically significant improvement in perceptual quality, with an LPIPS estimate of -0.023 ($t = -3.96$).

V. Discussion

The results presented in Section IV reveal a complex dynamic between environmental conditions and denoiser performance. While classical and neural methods share some behavioral trends, their divergence under extreme noise and varying conditions highlights fundamental differences in their operating mechanisms.

A. RQ1: Rate of performance degradation of Neural vs. Classical Methods

The comparative analysis addresses RQ1, comparing neural network-based denoisers (DRUNet) versus classical methods (BM3D) at varying noise levels. The ANOVA results (Table II) indicate a significant interaction between `Denoiser Type` and `Noise Level` for the LPIPS metric ($F = 19.25, p < 0.001$).

As shown in Fig. 2, while both methods degrade as noise increases, DRUNet exhibits a shallower degradation slope in perceptual quality (LPIPS) compared to BM3D. This can be attributed to the fundamental difference in their priors. BM3D relies on non-local self-similarity; as noise levels become extreme ($\sigma = 70$), the probability of finding similar patches decreases because the noise overwhelms the structural similarity, leading to artifacts. Conversely, DRUNet leverages learned deep priors. Even when the signal is severely corrupted, the

Fig. 3: Kodim01 image denoised with DRUNet under three different brightness settings: low, original, and high.

network can work out plausible high-frequency details based on its training data. While this may not always yield perfect pixel-alignment (hence the similar PSNR slopes), it preserves the realistic appearance of the image, resulting in significantly better LPIPS scores at high noise levels.

B. RQ2: The Divergence of PSNR and LPIPS

Addressing RQ2, the Linear Mixed Models (Table III) reveal a striking contradiction between signal fidelity (PSNR) and perceptual quality (LPIPS) regarding brightness and contrast. The models show that while high brightness and contrast significantly *improve* LPIPS (negative coefficient), they seemingly *degrade* PSNR (negative coefficient). This divergence arises from the way the different metrics penalize the loss of high-frequency detail versus the retention of noise variance. In low-brightness or low-contrast regions, the effective Signal-to-Noise Ratio (SNR) is poor. Under these conditions, both BM3D and DRUNet tend to aggressively smooth the image data to suppress noise. This over-smoothing can be visualized by comparing the bricks texture in Fig. 3.

- *Impact on PSNR:* Smoothing effectively removes noise variance. Even though textures and fine details are obliterated, the resulting pixel values settle near the local mean of the ground truth. Because PSNR is based on Mean Squared Error (MSE), it rewards this "safe bet." The reduction in variance mathematically outweighs the loss of texture, resulting in artificially inflated PSNR scores.
- *Impact on LPIPS:* Conversely, LPIPS relies on deep feature extraction to assess quality. The neural features interpret this over-smoothing as a complete loss of information. Consequently, LPIPS heavily penalizes this lack of texture.

Moreover, the improvement in LPIPS under high brightness aligns with the characteristics of human vision. The HVS (Human Visual System) is less sensitive to noise and artifacts in highly textured or bright areas (luminance masking). Since LPIPS is trained to mimic human perception, it "forgives" errors in bright regions, scoring them more favorably.

C. RQ3: Interaction of Contrast/Brightness and Noise Type

The analysis of interactions revealed distinct behaviors for brightness and contrast: brightness proved to be significantly more destructive on Poisson-degraded images than on Gaussian ones, from the perspective of both PSNR and LPIPS. Conversely, no definitive conclusions could be drawn regarding contrast interactions, given the low statistical significance of the factors and the conflicting trends between the two metrics.

VI. Conclusions

The results show that, at high noise levels, DRUNet achieves markedly better perceptual quality than BM3D, as captured by LPIPS, even when PSNR differences remain small. This highlights the limits of classical priors under severe corruption.

Brightness and contrast manipulations produced opposite trends in PSNR and LPIPS: higher brightness or contrast improved perceptual quality but reduced fidelity-based metrics, confirming that PSNR alone does not fully reflect visual realism under challenging photometric conditions.

Finally, interaction effects revealed that brightness and contrast enhancements benefit Gaussian-denoised images but have mixed impact on Poisson noise.

Furthermore, the reliance of neural denoisers like DRUNet on learned priors raises ethical concerns regarding data integrity. While these models effectively reconstruct plausible textures from severe noise, the divergence between improved perceptual scores (LPIPS) and signal fidelity (PSNR) suggests a risk of 'hallucinating' details that do not exist in the raw signal. In sensitive fields like healthcare and security, the choice of evaluation metric is crucial, as objective pixel fidelity should often take precedence over deep perceptual similarity. Additionally, the reliance on learned priors could create a risk of algorithmic bias. Future work must ensure that limited training diversity does not degrade image integrity for underrepresented subjects or textures.

References

[1] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark, 2018.

[2] Bhawna Goyal, Ayush Dogra, Sunil Agrawal, B.S. Sohi, and Apoorav Sharma. Image denoising review: From classical to state-of-the-art approaches. *Information Fusion*, 55:220–244, 2020.

[3] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65 vol. 2, 2005.

[4] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.

[5] 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.

[6] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2022.

[7] Francis J Anscombe. The transformation of poisson, binomial and negative-binomial data. *Biometrika*, 35(3/4):246–254, 1948.

[8] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.