# Is Nano Banana Pro a Low-Level Vision All-Rounder? A Comprehensive Evaluation on 14 Tasks and 40 Datasets

**Jialong Zuo**, **Haoyou Deng**, **Hanyu Zhou**, **Jiaxin Zhu**, **Yicheng Zhang**, **Yiwei Zhang**, **Yongxin Yan**, **Kaixing Huang**, **Weisen Chen**, **Yongtai Deng**, **Rui Jin**, **Nong Sang**, **Changxin Gao**

National Key Laboratory of Multispectral Information Intelligent Processing Technology,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology

The rapid evolution of text-to-image generation models has revolutionized visual content creation. While commercial products like Nano Banana Pro have garnered significant attention, their potential as generalist solvers for traditional low-level vision challenges remains largely underexplored. In this study, we investigate the critical question: **Is Nano Banana Pro a Low-Level Vision All-Rounder?** We conducted a comprehensive zero-shot evaluation across 14 distinct low-level tasks spanning 40 diverse datasets. By utilizing simple textual prompts without fine-tuning, we benchmarked Nano Banana Pro against state-of-the-art specialist models. Our extensive analysis reveals a distinct performance dichotomy: while **Nano Banana Pro demonstrates superior subjective visual quality**, often hallucinating plausible high-frequency details that surpass specialist models, **it lags behind in traditional reference-based quantitative metrics.** We attribute this discrepancy to the inherent stochasticity of generative models, which struggle to maintain the strict pixel-level consistency required by conventional metrics. This report identifies Nano Banana Pro as a capable zero-shot contender for low-level vision tasks, while highlighting that achieving the high fidelity of domain specialists remains a significant hurdle.

🍌 **Project Page:** https://lowlevelbanana.github.io

**GitHub:** https://github.com/zplusdragon/LowLevelBanana

🤗 **HuggingFace Dataset:** https://huggingface.co/datasets/jlongzuo/LowLevelEval

## 1 Introduction

The recent proliferation of Generative AI has fundamentally transformed the landscape of computer vision, with Text-to-Image (T2I) models demonstrating unprecedented capabilities in high-fidelity content creation. Among these, commercial products like Nano Banana Pro [138] have emerged as standouts, garnering significant attention for their versatility. While its prowess in creative synthesis is well-documented, the extent to which such a large-scale foundation model can generalize to traditional low-level vision problems remains largely unexplored. This gap presents not only a challenge of capability but also one of evaluation, raising the pivotal research question: **Is Nano Banana Pro a Low-Level Vision All-Rounder?**

The motivation for this study is rooted in a fundamental tension between human perception and traditional metrics. On one hand, the rich visual priors encapsulated within robust generative models should theoretically

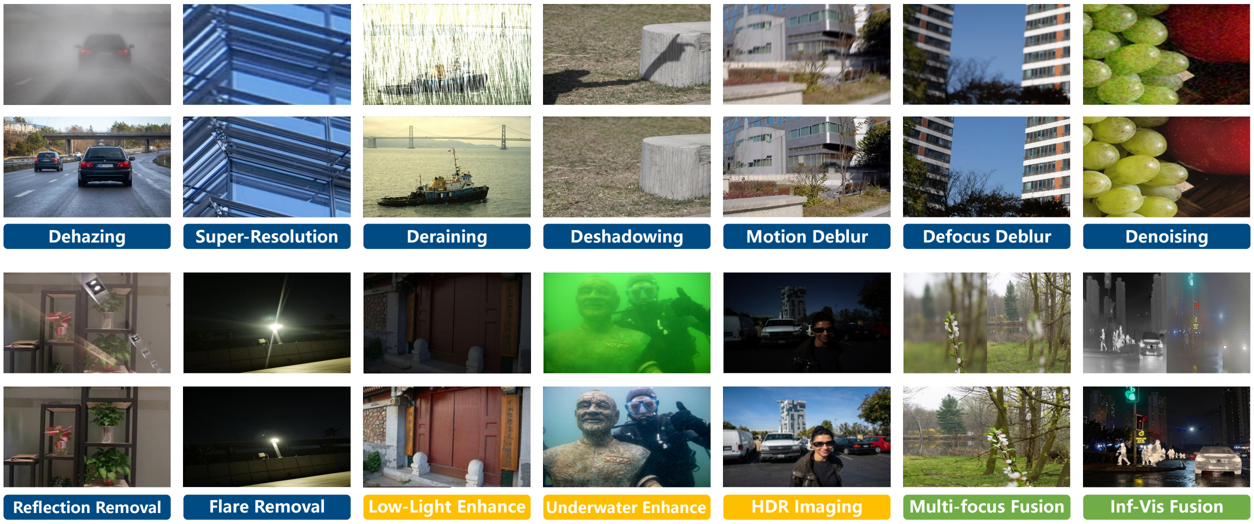


**Figure 1** Exemplary zero-shot results of Nano Banana Pro across 14 low-level vision tasks. For each task, the top row shows the degraded images, and the bottom row presents the corresponding restored outputs generated by Nano Banana Pro using simple text prompts. The visual results demonstrate the model's emerging capability for a diverse range of low-level vision tasks without task-specific training. The blue box represents image restoration tasks, the green box represents image enhancement tasks, and the yellow box represents image fusion tasks.

enable them to hallucinate plausible solutions for restoration, enhancement, and fusion tasks without task-specific training. On the other hand, this generative nature may conflict with the goal of achieving the strict, pixel-perfect fidelity that is valued by conventional evaluation metrics. To investigate this, we systematically evaluate the zero-shot capabilities of Nano Banana Pro using simple textual prompts, a stark contrast to the complex pipeline fine-tuning typically required for specialist models.

Our comprehensive empirical study spans 14 distinct low-level vision tasks across 40 datasets, covering image restoration, enhancement, and fusion. As visually exemplified in Fig. 1, Nano Banana Pro frequently produces outputs with remarkable perceptual quality. For instance, in tasks like dehazing or deraining, it can generate sharp edges and realistic textures that are often more aesthetically pleasing to a human observer than the results from specialist models. This initial observation immediately highlights a critical challenge: a model can be subjectively superior yet quantitatively inferior. Our work, therefore, aims not only to benchmark performance but also to delineate the boundaries of its current capabilities through rigorous quantitative and qualitative analysis. It is important to note that **the present evaluation reflects a conservative estimate of the model's capability**, as we did not engage in meticulous prompt tuning or employ multi-round inference to cherry-pick optimal outputs. Our fixed, simple prompts represent a pragmatic but unoptimized use case.

Our findings uncover this anticipated dichotomy in stark detail: **Nano Banana Pro excels in perceptual quality but lags in metric-driven fidelity.** While it demonstrates remarkable zero-shot potential across a wide array of degradations, its outputs consistently score lower on reference-based metrics (e.g., PSNR, SSIM) when compared to domain-specific experts. We attribute this performance gap to the inherent stochasticity of generative models, which prioritize semantic plausibility over the strict pixel-wise alignment demanded by these metrics. In essence, while Nano Banana Pro has not yet achieved the status of a perfect all-rounder, it forces us to reconsider the traditional definition of success in low-level vision. It establishes a strong baseline for zero-shot restoration, highlighting both its emerging strengths and the critical need for new evaluation paradigms that can reconcile perceptual quality with pixel-level accuracy.

The remainder of this report is organized to systematically present these findings. We will detail our experimental results across Image Restoration, Image Enhancement, and Image Fusion, providing in-depth comparative analysis for each task. Finally, we conclude by summarizing the model's limitations and discussing potential future directions, including the development of more perception-aligned evaluation methods for generative low-level vision solvers.

## Contents

# Image Restoration

## 2 Dehazing

### 2.1 Introduction

Real-world Image Dehazing (RID) aims to recover clear and high-fidelity images from hazy observations captured in real-world environments. Unlike synthetic dehazing, where haze is generated under simplified physical assumptions, real-world haze is highly complex and exhibits strong spatial non-uniformity, large depth variations, severe color shifts, sensor noise, and compression artifacts. These factors make RID a long-standing and extremely challenging low-level vision problem. The goal of RID is not only to generate visually appealing results, but also to preserve accurate photometric and structural information to support reliable downstream vision tasks such as detection, tracking, and segmentation.

Early dehazing methods [119, 134] relied on handcrafted statistical priors, such as the Dark Channel Prior (DCP)[50] and Non-Local Prior (NLP)[8], to constrain the solution space. While these physics-inspired approaches achieved initial success, they often fail to generalize across diverse real-world scenes and frequently introduce visible artifacts. With the rapid development of deep learning, numerous CNN-based and Transformer-based methods[19, 131, 163], have significantly improved dehazing performance on synthetic benchmarks. However, collecting large-scale, perfectly aligned real-world hazy and clean image pairs remains nearly impossible. Although several real-world datasets have been constructed, their scale and diversity are still far from sufficient for training robust deep models. Consequently, most existing methods heavily rely on synthetic data and suffer from severe performance degradation when deployed in real-world scenarios.

To bridge this gap, recent studies have increasingly shifted their focus toward real-world dehazing. Some methods reintroduce physical priors to adapt pre-trained networks, while others modify inference strategies to improve generalization. Nevertheless, these approaches remain highly dependent on the quality of pre-training data. Moreover, heavily hazed images often contain severe information loss, fundamentally limiting the capability of traditional enhancement-based methods that lack generative flexibility to recover missing content.

Nano Banana is an image generation model developed by Google DeepMind. Its professional version, Nano Banana Pro, further enhances precision and world knowledge understanding. We applied it to real-world image dehazing tasks, with a focus on its effectiveness in removing haze, restoring blurred textures, and maintaining scene semantic integrity and tested it on mainstream dehazing benchmarks.

### 2.2 Quantitative and Qualitative Results

**Table 1** Quantitative comparison of dehazing methods on multiple datasets. NB Pro achieves excellent NIMA scores on both datasets. It also demonstrates favorable FADE and BRISQUE scores on the RTTS dataset. However, its FADE and BRISQUE metrics on the Fattal's dataset are unsatisfactory. The best results are in **black bold.**

| Method | RTTS | | | Fattal's | | |
|---|---|---|---|---|---|---|
| | FADE↓ | BRISQUE↓ | NIMA↑ | FADE↓ | BRISQUE↓ | NIMA↑ |
| MBSDN [29] | 1.363 | 27.67 | 4.53 | 0.579 | **14.15** | 5.43 |
| Dehamer [43] | 1.895 | 33.24 | 4.52 | 0.698 | 15.53 | 5.16 |
| DAD [131] | 1.130 | 32.24 | 4.31 | 0.484 | 29.64 | **5.46** |
| PSD [19] | 0.920 | 27.71 | 4.60 | 0.416 | 23.61 | 4.99 |
| D4[180] | 1.358 | 33.21 | 4.48 | 0.411 | 20.33 | 5.44 |
| RIDCP[163] | 0.944 | 17.29 | 4.97 | 0.408 | 20.05 | 5.43 |
| CORUN [32] | **0.824** | **11.96** | **5.34** | **0.338** | 14.82 | 5.39 |
| **NB Pro** | **0.986** | **27.21** | **4.95** | **0.683** | **22.16** | **5.44** |

To intuitively present the dehazing outcomes of the Nano Banana (NB) Pro model, we provide quantitative and qualitative evaluations of its processing results across the RTTS [74] and Fattal's [33] datasets, with comparisons to state-of-the-art baseline methods in real-world dehazing tasks. Tab. 1 show the performance comparison between NB Pro and current mainstream dehazing networks on the RTTS and Fattal's datasets, where we evaluated three real-world-oriented dehazing metrics: FADE , BRISQUE and NIMA. Integrating the evaluation results on both the RTTS and Fattal's datasets, NB Pro demonstrates outstanding performance in terms of subjective visual quality, achieving top-tier NIMA scores on both benchmarks, which indicates that the generated images possess strong aesthetic appeal and favorable human perceptual quality. However, it performs poorly in terms of image naturalness. Specifically, NB Pro exhibits a significantly higher BRISQUE score on the Fattal's dataset, suggesting that the outputs may suffer from over-enhancement artifacts.

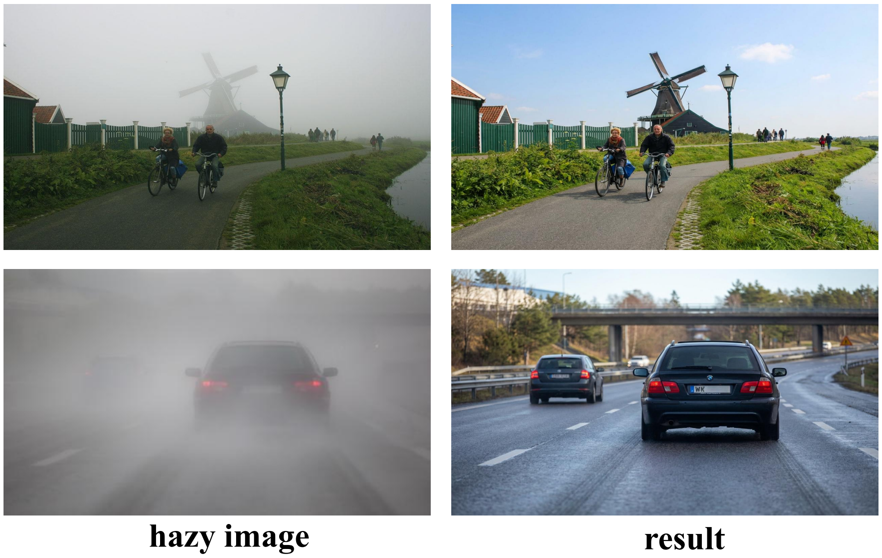


**Figure 2** NB Pro dehazing visual results on the RTTS dataset. Especially under heavy hazy conditions, NB Pro can effectively recover and enhance blurred details.

Qualitative experimental results demonstrate that for extreme degradation scenarios such as dense fog with severe visibility loss, heavy atmospheric scattering, and complex urban or natural environments, NB Pro can generate perceptually enhanced results. Fig. 2 displays the visualization of exemplary dehazing cases for NB Pro on the RTTS dataset, highlighting both successful and challenging scenarios. For certain severely hazy images, benefiting from its powerful generative capabilities, NB Pro effectively restores intricate details—such as fine textures in buildings, vehicles, or vegetation—that are heavily obscured in the input, producing clear, high contrast outputs with impressive visual recovery of distant structures and overall scene coherence. Similarly, for some lightly foggy images, NB Pro performs well in optimization, selectively removing haze from distant backgrounds while preserving foreground sharpness and natural tones, resulting in balanced enhancements that improve visibility without introducing artifacts. Fig. 3 shows the visual comparison of NB and other methods on the RTTS dataset.

However, NB Pro also exhibits numerous failure cases, as illustrated in Fig. 4. In these examples, spanning both moderate and heavy haze conditions, NB Pro often restores the images into distorted outputs characterized by unnatural color shifts, such as over-saturated or overly vibrant hues and hallucinatory weather elements, most notably forcing intensely blue skies into scenes that were originally overcast or neutral. These distortions undermine the authenticity of the atmospheric conditions, leading to results that deviate significantly from realistic dehazing expectations.

Overall, while NB Pro offers inspirational generative capabilities for ill-posed real-world dehazing, demonstrating the potential of semantic-driven priors in tackling highly ambiguous degradations—its limitations in color fidelity, physical realism, and consistent handling across varying haze densities suggest it is better suited for creative enhancements rather than precise low-level restoration. This prompts future research into hybrid approaches, such as combining NB Pro's zero-shot generative strengths with task-specific physical constraints or refined prompt engineering, to better constrain its tendencies toward perceptual appeal over faithful recovery.

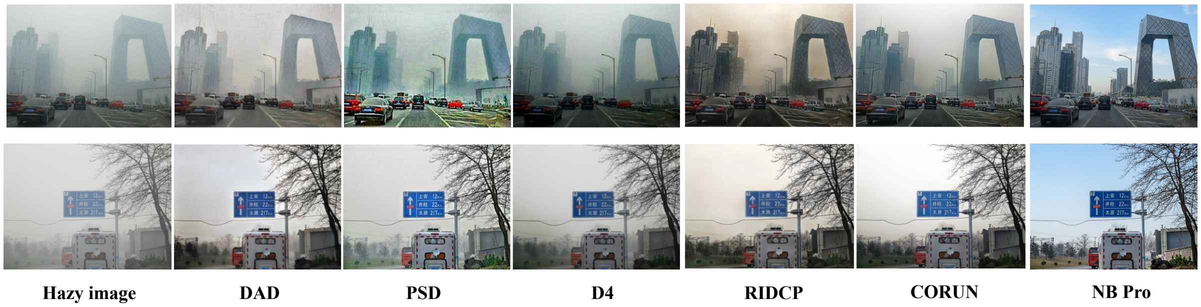


**Figure 3** A comparison of visual results between NB pro and other methods. It can be observed that the results produced by NB Pro are noticeably clearer and exhibit superior visual quality; however, they also suffer from obvious over-enhancement artifacts.

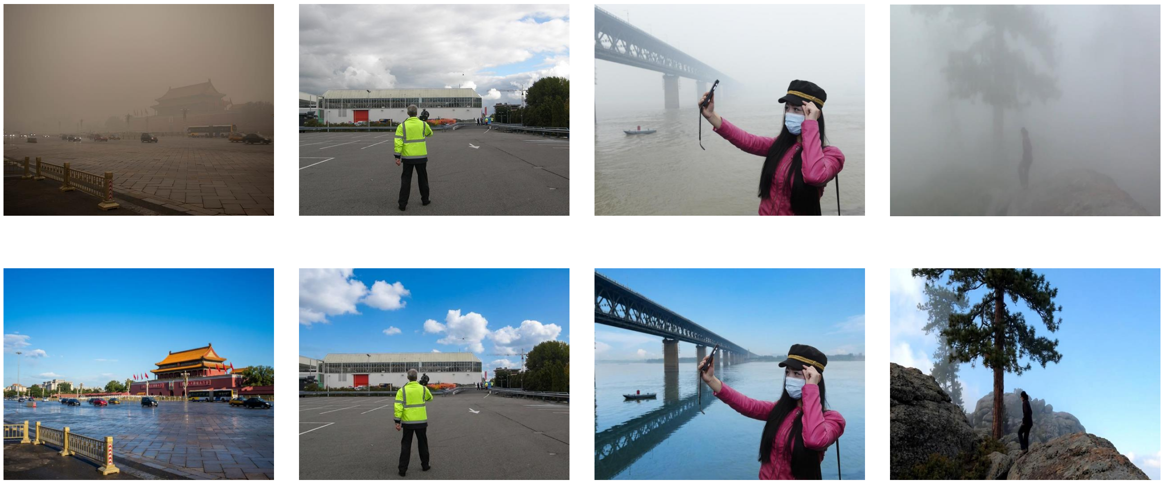


**Figure 4** Anamorphic example images of Nano Pro in dehazing on the RTTS dataset. The original hazy images is on top, and below is the results after hazy removal. The dehazed results exhibit poor color fidelity.

## 2.3 Analyses

Stemming from a misalignment between training distributions and restoration objectives, NB Pro struggles to maintain spectral fidelity and atmospheric consistency. The model frequently over-corrects naturally muted, hazy tones into saturated, vibrant colors, introducing artificial chromatic biases that alter the scene's intrinsic mood. In severely hazy scenarios where high-frequency details are obliterated, NB Pro's generative priors dominate the restoration process. Rather than solving the inverse physical scattering model, it hallucinates details based on learned statistical patterns. A defining characteristic of this behavior is the systematic rendering of vivid blue skies in originally overcast scenes. While visually striking and aligned with subjective preferences for "clear weather", this deviation undermines the scene's authenticity and temporal consistency.

Consequently, NB Pro is currently best positioned for creative content generation—prioritizing perceptual appeal and "wow-factor"—rather than forensic restoration tasks demanding strict adherence to physical constraints. However, its value remains significant. In the ill-posed domain of real-world dehazing, where traditional physics-based methods often falter due to unknown degradation parameters, NB Pro demonstrates the power of semantic priors to reconstruct plausible details in information-deficient scenarios. This suggests a pivotal direction for future research: developing hybrid frameworks. By integrating NB Pro-like generative backbones with task-specific physical constraints (e.g., atmospheric scattering laws) and fidelity anchors, we can bridge the gap between creative hallucination and realistic restoration, aiming for a paradigm that balances visual delight with physical truth.

# 3 Super-Resolution

## 3.1 Introduction

Real-World Image Super-Resolution (Real-ISR) aims to restore high-fidelity, high-resolution content from low-resolution inputs degraded by complex, unknown physical processes. Unlike synthetic super-resolution, where degradations are modeled by simple bicubic downsampling, real-world scenarios involve intricate combinations of blur, sensor noise, compression artifacts, and varying camera response functions [12, 162]. This complexity renders traditional regression-based methods—which rely on pixel-wise optimization (e.g., MSE loss)—ineffective, often resulting in over-smoothed textures and a lack of high-frequency details [28, 201].

The landscape of generative Real-ISR has evolved rapidly. Generative Adversarial Networks (GANs) [42, 72], represented by methods such as BSRGAN [193] and Real-ESRGAN [152], introduced high-order degradation modeling to synthesize realistic training pairs, significantly improving visual perceptual quality over PSNR-oriented baselines. More recently, Denoising Diffusion Probabilistic Models (DDPMs) [51] have emerged as the new state-of-the-art. Methods like StableSR [146] and DiffBIR [93] leverage strong priors from large-scale pre-trained text-to-image models (e.g., Stable Diffusion [128]) to generate intricate textures. However, these multi-step diffusion models often suffer from high computational costs and slow inference speeds. To mitigate this, acceleration techniques have been proposed, exemplified by SinSR [155], which distills complex diffusion priors into single-step inference models. Despite these advancements, a critical challenge remains: the inherent trade-off between perceptual quality and signal fidelity [9], often leading to artifacts or structural hallucinations in the pursuit of sharpness.

In this work, we conduct a comprehensive quantitative and qualitative evaluation of Nano Banana Pro, a novel generative ISR framework, benchmarking it against a spectrum of industry-standard algorithms, including GAN-based approaches (BSRGAN [193], Real-ESRGAN [152]), multi-step diffusion models (StableSR [146], DiffBIR [93]), and accelerated diffusion methods (SinSR [155]). Our goal is to rigorously assess where Nano Banana Pro stands within the current Perception-Distortion landscape.

To ensure a robust evaluation across varying degradation complexities, our experiments are conducted on the large-scale DIV2K-Val dataset (2,994 images) [4] as well as the authentic RealSR [12] and DRealSR [162] benchmarks. Recognizing that pixel fidelity alone is insufficient for characterizing generative performance, we employ a comprehensive set of evaluation metrics, incorporating not only standard Full-Reference indicators (PSNR, SSIM [159], LPIPS [196]) but also widely-adopted No-Reference perceptual metrics (NIQE [112], MUSIQ [69], CLIPIQA [145]). Under this rigorous testing framework, we systematically assess the reconstruction capabilities of Nano Banana Pro in comparison to established GAN-based and diffusion-based baselines. The resulting analysis provides a detailed characterization of the model's behavior regarding the perception-distortion trade-off, offering valuable insights into its suitability for real-world applications.

## 3.2 Quantitative Results

To comprehensively evaluate Nano Banana Pro's performance in Real-ISR tasks, we quantitatively compared it against a range of advanced GAN-based and diffusion-based image super-resolution methods. We employed standard full-reference metrics: PSNR and SSIM to evaluate signal fidelity, and LPIPS to assess perceptual similarity. Additionally, No-Reference (NR) metrics NIQE, MUSIQ, and CLIPIQA were utilized to quantify the statistical naturalness and aesthetic quality of the generated images. Results are shown in Tab. 2. Nano Banana Pro significantly underperformed against the comparison methods in terms of traditional fidelity metrics. On the DIV2K-Val dataset, Nano Banana Pro achieved significantly lower PSNR and SSIM than the optimal method, lagging by over 4 dB. A similar trend, though less severe, was observed on the RealSR and DRealSR datasets, where its fidelity scores remained consistently behind both GAN-based and diffusion-based baselines. This result clearly indicates that under the standard full-reference evaluation framework, which prioritizes pixel-level accurate reconstruction, Nano Banana Pro's generated results exhibit systematic deviations from the ground-truth reference images.

Nano Banana Pro fundamentally differs from traditional super-resolution models optimized for specific degradation kernels. The latter typically undergo end-to-end training targeting minimization of pixel-level loss—thus inherently excelling in metrics like PSNR and SSIM. In contrast, Nano Banana Pro's

**Table 2** Quantitative comparison on synthetic (DIV2K-Val[4]) and real-world (RealSR[12], DRealSR[162]) benchmarks. The best and second best results are highlighted by **black bold** and underline.

| Test Dataset | Method | Full-Reference Metrics | | | No-Reference Metrics | | |
|---|---|---|---|---|---|---|---|
| | | PSNR↑ | SSIM↑ | LPIPS↓ | NIQE↓ | MUSIQ↑ | CLIPIQA↑ |
| DIV2K-Val | BSRGAN [193] | **24.58** | 0.6269 | 0.3351 | 4.75 | 61.20 | 0.5071 |
| | Real-ESRGAN [152] | 24.29 | **0.6371** | **0.3112** | 4.68 | 61.06 | 0.5501 |
| | StableSR [146] | 23.26 | 0.5726 | 0.3113 | 4.76 | **65.92** | 0.6192 |
| | DiffBIR [93] | 23.64 | 0.5647 | 0.3524 | 4.70 | 65.81 | **0.6210** |
| | SinSR [155] | 24.41 | 0.6018 | 0.3240 | 6.02 | 62.82 | 0.5386 |
| | Nano Banana Pro | 20.29 | 0.4720 | 0.3645 | **3.52** | 65.40 | 0.5257 |
| RealSR | BSRGAN | **26.39** | **0.7654** | **0.2670** | 5.66 | 63.21 | 0.5001 |
| | Real-ESRGAN | 25.69 | 0.7616 | 0.2727 | 5.83 | 60.18 | 0.4449 |
| | StableSR | 24.70 | 0.7085 | 0.3018 | 5.91 | **65.78** | 0.6221 |
| | DiffBIR | 24.75 | 0.6567 | 0.3636 | 5.53 | 64.98 | **0.6246** |
| | SinSR | 26.28 | 0.7347 | 0.3188 | 6.29 | 60.80 | 0.5385 |
| | Nano Banana Pro | 23.56 | 0.6649 | 0.2978 | **4.39** | 60.18 | 0.5199 |
| DRealSR | BSRGAN | **28.75** | 0.8031 | 0.2883 | 6.52 | 57.14 | 0.4915 |
| | Real-ESRGAN | 28.64 | **0.8053** | **0.2847** | 6.69 | 54.18 | 0.4422 |
| | StableSR | 28.03 | 0.7536 | 0.3284 | 6.52 | 58.51 | 0.5601 |
| | DiffBIR | 26.71 | 0.6571 | 0.4557 | 6.31 | **61.07** | **0.5930** |
| | SinSR | 28.36 | 0.7515 | 0.3665 | 6.99 | 55.33 | 0.4884 |
| | Nano Banana Pro | 23.97 | 0.6323 | 0.3809 | **5.03** | 59.00 | 0.5145 |

generation process prioritizes semantic coherence and visual cleanliness. Its outputs can be viewed as plausible reconstructions of the input image rather than strict pixel-to-pixel mappings. Consequently, generated images may exhibit deviations in local texture alignment and structural positioning compared to reference images, leading to comprehensive score reductions across full-reference metrics. However, notably, on the NIQE metric, Nano Banana Pro consistently achieved the best scores across all three datasets (e.g., 3.52 on DIV2K-Val vs. 4.75 for BSRGAN). This suggests its outputs possess superior statistical naturalness, effectively suppressing artifacts, even though the low-level pixel arrangements have been altered from the original reference.

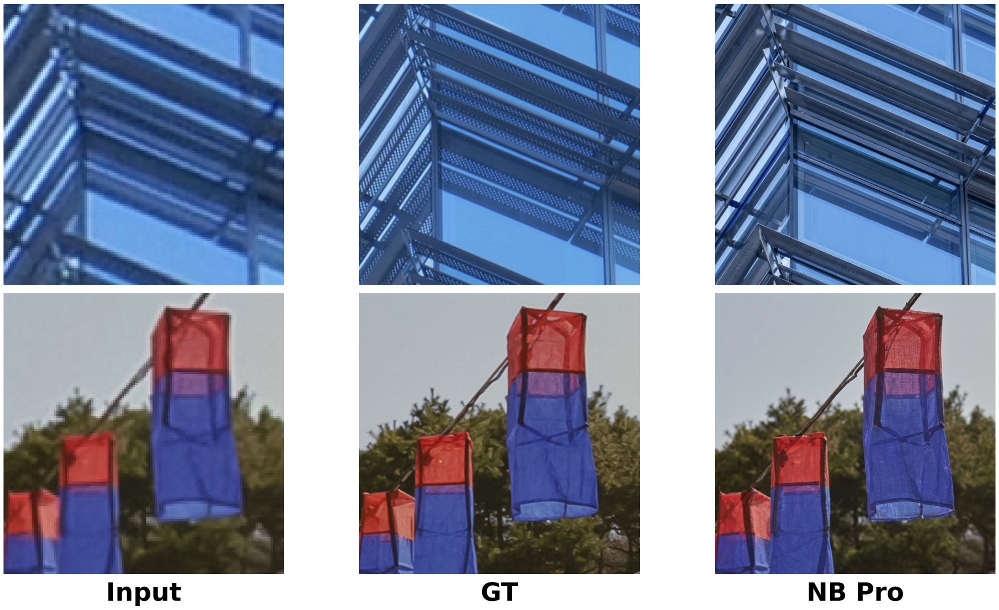


**Figure 5** Qualitative visualization of structural recovery in Real-ISR tasks using Nano Banana Pro.

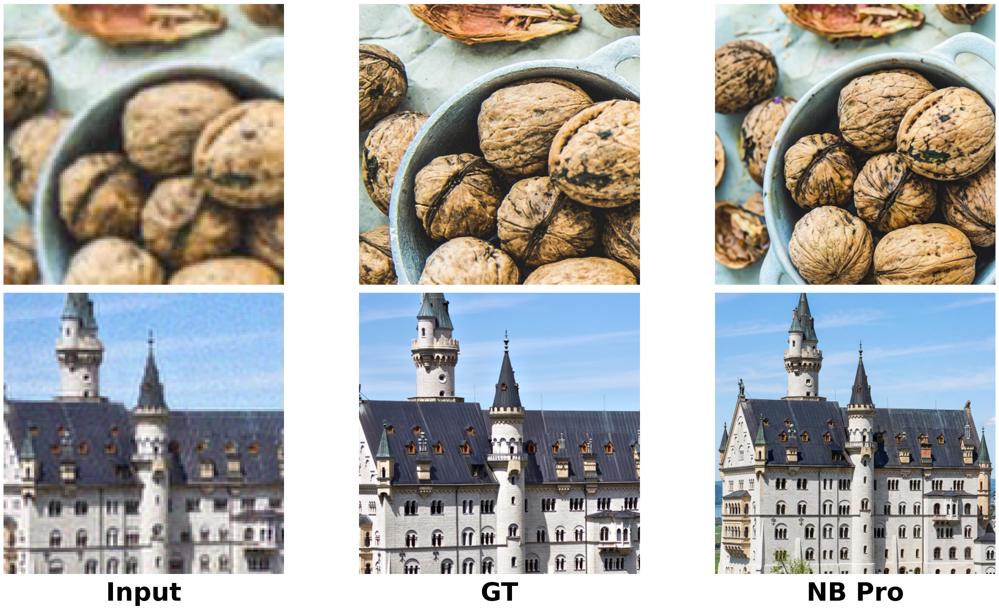


**Figure 6** Visualization of unintended image boundary extension cases in Real-ISR tasks using Nano Banana Pro.

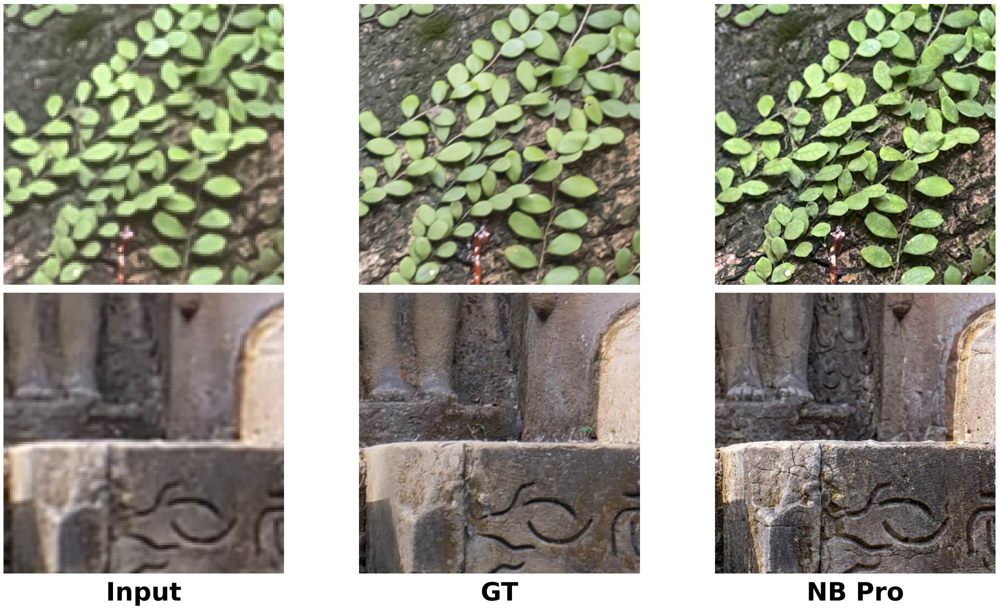


**Figure 7** Visualization of generative texture deviations in Real-ISR tasks using Nano Banana Pro.

## 3.3 Qualitative Results

In this section, we examine the generative characteristics of Nano Banana Pro across the DIV2K, RealSR, and DRealSR datasets. The qualitative evaluation is organized into four key scenarios to highlight both the strengths and failure modes of the model: geometric clarity, field-of-view artifacts, texture fidelity, and character reconstruction (Figs. 5–8).

Fig. 5 displays the super-resolution results on scenes with distinct geometric structures. In the examples of the architectural facade and hanging lanterns, Nano Banana Pro effectively sharpens the blurred edges and recovers the linear patterns lost in the low-resolution inputs. The resulting images maintain structural coherence and exhibit reduced noise, offering a noticeable improvement in clarity compared to the inputs.

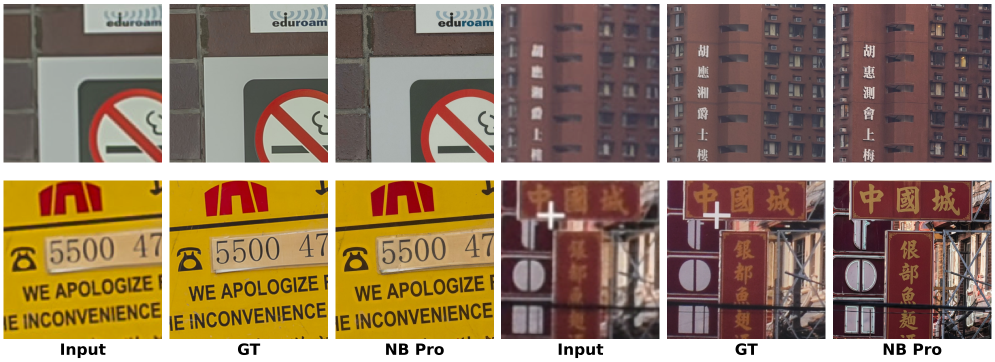


**Figure 8** This visualization illustrates text reconstruction, showing successful character recovery in the left panel and erroneous character hallucination in the right panel.

Fig. 6 illustrates a distinct structural anomaly observed in Nano Banana Pro: unintended Field-of-View (FOV) expansion. Comparing the low-resolution input, the Ground Truth (GT), and the Nano Banana Pro generated result reveals that the model fails to strictly adhere to the original spatial boundaries of the input image. Instead of merely super-resolving the existing pixels, Nano Banana Pro erroneously hallucinates additional content along the image periphery.

Fig. 7 illustrates the tendency of Nano Banana Pro to synthesize plausible but non-existent details in areas with complex stochastic textures. In the foliage and stone carving examples, while the model produces sharp, high-frequency patterns, these generated textures often deviate from the Ground Truth. Specifically, the arrangement of leaf veins and the granular surface of the stone are reconstructed with altered local structures rather than being faithfully restored, leading to pixel-level discrepancies that lower fidelity scores.

Fig. 8 highlights the dependency of Nano Banana Pro on semantic recognizability for text reconstruction. In the examples on the left, where the low-resolution input retains discernible structural features (such as the logo and digits), the model accurately reconstructs sharp and legible characters. Conversely, the examples on the right demonstrate failure cases where severe degradation obscures the original glyphs, particularly with complex Chinese characters. In these instances, the model fails to recover the correct semantic content and instead hallucinates incorrect strokes or non-existent characters, resulting in high-contrast but semantically erroneous text.

## 3.4 Analyses

Our comprehensive evaluation elucidates the operational characteristics of Nano Banana Pro within the Real-ISR domain. Quantitatively, the model trails significantly behind mainstream GAN and diffusion-based methods in fidelity metrics (PSNR/SSIM) across the DIV2K-Val and RealSR datasets; however, it achieves remarkable performance in the no-reference NIQE metric. This discrepancy suggests that, as evidenced in Fig. 7, the model prioritizes learned generative priors to synthesize texture details rather than strictly adhering to the low-resolution input.

Qualitatively, unlike traditional restoration models that maintain strict spatial consistency, Nano Banana Pro lacks precise pixel-level alignment with the reference image. Consequently, unintended Field-of-View (FOV) expansion is observed, as shown in Fig. 6. Furthermore, the text reconstruction failures in Fig. 8 reveal the model's heavy reliance on feature recognizability. When encountering degraded features such as blurred text, the model exhibits a tendency to aggressively generate sharp outputs. This behavior causes feature recognition errors to have a catastrophic impact on the result, leading to the hallucination of sharp but semantically incorrect text.

Synthesizing these findings, Nano Banana Pro is highly suitable for perception-centric applications—such as artistic upscaling, old photo restoration, or casual photography—where visual purity and noise elimination are paramount. However, due to its propensities for texture hallucination, spatial misalignment, and semantic alteration, it is unsuitable for fidelity-critical scenarios.

# 4 Deraining

## 4.1 Background

Rain is a common yet challenging weather degradation that severely obscures scene content and alters the structural appearance of images. Such degradations significantly reduce visual quality and adversely affect the reliability of numerous outdoor vision systems, including intelligent transportation, UAV-based monitoring, and autonomous navigation. Single image deraining, which seeks to restore a clean background from a rain-contaminated observation, has therefore become an essential task in low-level vision. In recent years, a variety of deraining algorithms and benchmark datasets have been developed, achieving remarkable progress in modeling rain streaks, raindrops, and atmospheric veils.

Recent advances in large-scale multimodal generative models offer a promising new direction for image restoration. Among these models, Nano Banana Pro, Google's latest high-fidelity visual generation system built upon the powerful Gemini 3 Pro multimodal reasoning engine, excels in semantic comprehension, fine-grained visual modeling, and precise structural control. Designed for professional image creation and editing, Nano Banana Pro supports high-resolution synthesis, multi-image fusion, accurate text rendering, and semantically coherent scene manipulation. These capabilities suggest that the model is not only adept at generating novel visual content, but also inherently possesses strong prior knowledge for reconstructing clean structures and textures, rendering it a promising candidate for image restoration tasks such as deraining.

In this study, we investigate the feasibility of adapting Nano Banana Pro to single-image deraining. In contrast to traditional restoration networks that rely on explicit task-specific modeling, Nano Banana Pro utilizes its rich world model and multimodal reasoning ability to interpret degraded regions, infer plausible background structures, and produce natural, artifact-free reconstructions. By framing deraining as a guided generative reconstruction problem, we aim to harness the model's semantic priors and sophisticated visual synthesis capabilities to achieve rain removal across diverse scenarios.

## 4.2 Experiment Setup

To thoroughly evaluate the performance and generalization ability of Nano Banana Pro on the single image deraining task, we conduct experiments on three widely used benchmark datasets: two synthetic datasets, Rain200L and Rain200H[178], and one real-world dataset, SPA[150]. These datasets cover a broad range of rain patterns and scene complexities, enabling a comprehensive assessment of the model's restoration capability.

- **Rain200L**[178]: Contains 1,800 pairs of synthetic training images and 200 test pairs. The rain streaks in this dataset exhibit a single predominant direction and relatively low density, making it suitable for assessing the model's basic restoration ability under simple rain conditions.
- **Rain200H**[178]: Provides 1,800 training pairs and 200 test pairs, but features rain streaks with higher density and more diverse orientations. This dataset is designed to evaluate the robustness of deraining models when faced with heavy and structurally complex rain degradations.
- **SPA**[150]: A large-scale real-world rainy image dataset comprising 638,492 training pairs and 1,000 test images. The rain patterns in SPA are highly diverse, and the background scenes vary significantly, making it an appropriate benchmark for measuring cross-domain generalization from synthetic to real rainy conditions.

All images from all datasets are given the prompt: *"This is a rainy image. Please remove the rain streaks and raindrops while keeping all other elements, the original color tone, lighting, and atmosphere unchanged."*

It is worth emphasizing that Nano Banana Pro is evaluated in a strictly zero-shot manner: no training images are used for optimization, fine-tuning, or adaptation, and the model is directly applied to the test images via a fixed textual prompt.

### 4.3 Quantitative and Qualitative Results

For fair comparison, all metrics are computed under exactly the same evaluation protocol as NeRD-Rain, including image resolution, color space, and PSNR/SSIM computation. Based on the quantitative results reported in Tab. 3, we systematically evaluate the image deraining performance of Nano Banana Pro on two synthetic datasets, Rain200L[178] and Rain200H[178], as well as the real-world SPA-Data dataset[150], and compare it with a wide range of representative methods. The compared approaches cover prior-based methods (DSC[103], GMM[90]), CNN-based methods (DDN[35], RESCAN[87], PReNet[126], MSPFN[67], RCDNet[144], MPRNet[183], DualGCN[36], SPDNet[181]), and Transformer-based methods (Uformer[157], Restormer[184], IDT[167], DRSformer[16], NeRD-Rain[17]).

**Table 3** Quantitative comparison results on three representative benchmarks. The best results are in **black bold.**

| Method | Rain200L | | Rain200H | | SPA-Data | |
|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| DSC[103] | 27.16 | 0.8663 | 14.73 | 0.3815 | 34.95 | 0.9416 |
| GMM[90] | 28.66 | 0.8652 | 14.50 | 0.4164 | 34.30 | 0.9428 |
| DDN[35] | 34.68 | 0.9671 | 26.05 | 0.8056 | 36.16 | 0.9457 |
| RESCAN[87] | 36.09 | 0.9697 | 26.75 | 0.8353 | 38.11 | 0.9707 |
| PReNet[126] | 37.80 | 0.9814 | 29.04 | 0.8991 | 40.16 | 0.9816 |
| MSPFN[67] | 38.58 | 0.9827 | 29.36 | 0.9034 | 43.43 | 0.9843 |
| RCDNet[144] | 39.17 | 0.9885 | 30.24 | 0.9048 | 43.36 | 0.9831 |
| MPRNet[183] | 39.47 | 0.9825 | 30.67 | 0.9110 | 43.64 | 0.9844 |
| DualGCN[36] | 40.73 | 0.9886 | 31.15 | 0.9125 | 44.18 | 0.9902 |
| SPDNet[181] | 40.50 | 0.9875 | 31.28 | 0.9207 | 43.20 | 0.9871 |
| Uformer[157] | 40.20 | 0.9860 | 30.80 | 0.9105 | 46.13 | 0.9913 |
| Restormer[184] | 40.99 | 0.9890 | 32.00 | 0.9329 | 47.98 | 0.9921 |
| IDT[167] | 40.74 | 0.9884 | 32.10 | 0.9344 | 47.35 | 0.9930 |
| DRSformer[16] | 41.23 | 0.9894 | 32.17 | 0.9326 | 48.54 | 0.9924 |
| NeRD-Rain[17] | **41.71** | **0.9903** | **32.40** | **0.9373** | **49.58** | **0.9940** |
| **Nano Banana Pro** | **26.05** | **0.7954** | **21.10** | **0.6659** | **32.25** | **0.9142** |

The quantitative performance of Nano Banana Pro is significantly inferior to that of state-of-the-art deraining models across all three datasets. On Rain200L[178], Nano Banana Pro achieves 26.05 dB PSNR and 0.7954 SSIM, which are substantially lower than those of supervised learning-based methods. On the more challenging Rain200H dataset with complex rain patterns, Nano Banana Pro obtains 21.10 dB PSNR and 0.6659 SSIM. Although it outperforms traditional prior-based methods, a considerable performance gap remains compared to CNN-based and Transformer-based approaches. On the real-world SPA-Data dataset[150], Nano Banana Pro reaches 32.25 dB PSNR and 0.9142 SSIM, still noticeably below the current best-performing methods.

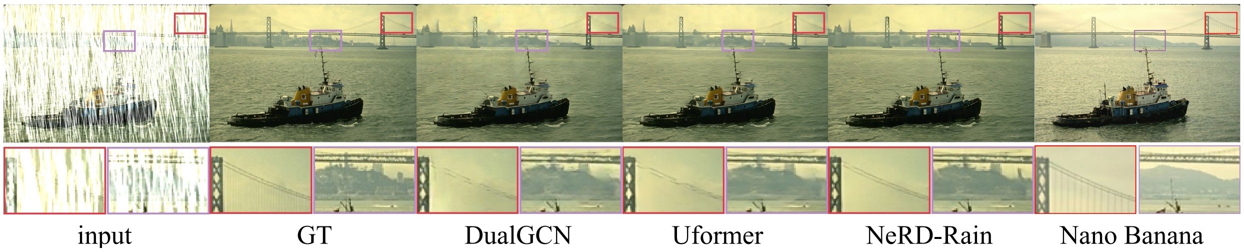


**Figure 9** Qualitative comparison on the Rain200H[178] dataset. Close-up views better illustrate the deraining capability.

This quantitative degradation is consistent with the visual results in Fig. 9, where large and dense synthetic rain streaks severely obscure background content, leading to color deviations and missing or oversmoothed fine details. Since Nano Banana Pro is not trained specifically for image deraining, local structures are often reconstructed via implicit generative hallucination rather than pixel-wise restoration, which negatively affects PSNR and SSIM. Nevertheless, the model demonstrates notable strength in recovering certain global structures; for example, the cable geometry of the suspension bridge is reconstructed with higher structural coherence and semantic plausibility than several supervised baselines.

Furthermore, Fig. 10 shows that the deraining performance of Nano Banana Pro is highly sensitive to rainfall

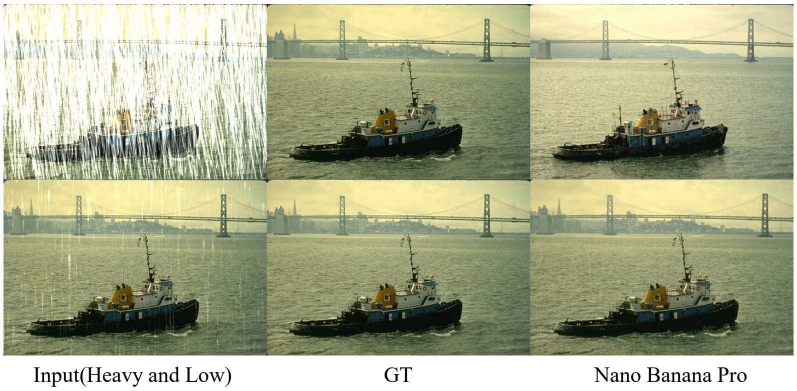


**Figure 10** Qualitative comparison under different rain intensities. Under lighter rain, Nano Banana Pro better preserves the original tone and fine details, while heavier rain leads to noticeable color shifts and detail loss.

intensity: under low-rain conditions, where more reliable visual information is preserved, both color fidelity and fine details are significantly improved, whereas heavy rainfall introduces severe occlusion and ambiguity, resulting in pronounced color shifts and detail degradation. Overall, these results indicate that while generative multimodal models are disadvantaged in pixel-level fidelity under zero-shot deraining, they retain strong semantic and structural priors, suggesting complementary potential in scenarios with limited supervision or severe information loss.

Our qualitative analysis reveals a fundamental limitation in the instruction-following behavior of prompt-conditioned multimodal generative models. As shown in Fig. 11, despite utilizing explicit prompts that constrain the model to preserve non-rain regions, we consistently observe unintended alterations in background elements.

This phenomenon stems from an inherent bias in fine-grained semantic understanding. As clearly illustrated in Fig. 12, the model conflates rain streaks with atmospheric haze (i.e., the rain-mist ambiguity). Consequently, it aggressively removes the mist alongside the rain, yielding an output image that appears clearer and visually superior in low-level details. However, this visual enhancement paradoxically leads to lower quantitative scores (e.g., PSNR), as the complete removal of haze introduces a significant pixel-wise deviation from the ground truth. This observation underscores the prevalent perception-distortion trade-off in image restoration tasks.

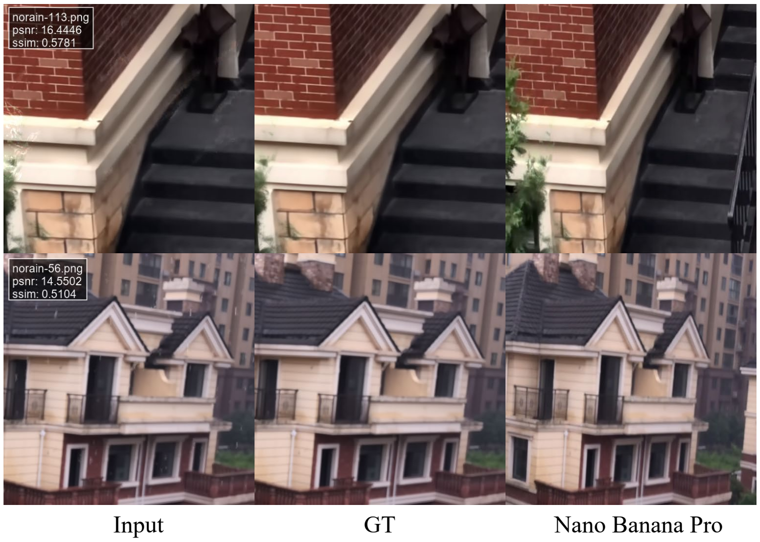


**Figure 11** Failure case of Nano Banana Pro. From left to right: input rainy image, ground truth (GT), and the output of Nano Banana Pro. The model hallucinates plausible content to fill severely corrupted regions, leading to a significant pixel-level discrepancy from the GT.

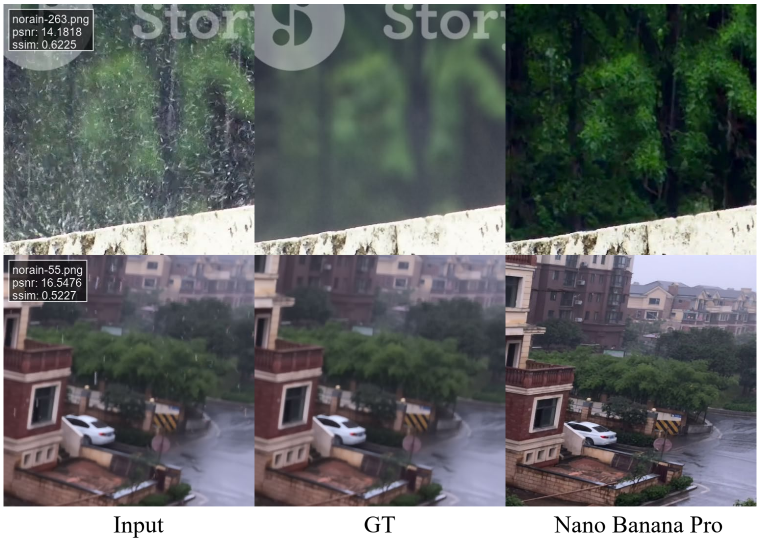


**Figure 12** Failure case of Nano Banana Pro. The model inadvertently removes the atmospheric haze accompanying rain streaks; while this enhances visual clarity, it results in lower quantitative metrics due to deviation from the GT.

### 4.4 Analyses

In this study, we investigated the feasibility of Nano Banana Pro for single-image deraining under a zero-shot setting. Experimental results indicate that the application of generative models to image restoration presents a “double-edged sword” effect. On one hand, compared to specialized deraining models trained on extensive supervised data, such as NeRD-Rain and Restormer, Nano Banana Pro exhibits a significant gap in pixel-level objective metrics like PSNR and SSIM (e.g., achieving a PSNR of only 21.10 dB on the Rain200H dataset). This quantitative deficiency is primarily attributed to the inherent tendency of generative models to prioritize semantic reconstruction over strict pixel-wise restoration, resulting in deviations in high-frequency detail preservation and color fidelity. On the other hand, leveraging its robust world model and semantic reasoning capabilities, Nano Banana Pro demonstrates superior structural coherence and visual plausibility compared to traditional methods when handling regions with severe rain occlusion (such as bridge cables). This confirms the unique complementary advantages of generative models in addressing extreme degradation and information loss. Future work will focus on employing prompt tuning to further enhance pixel-level restoration accuracy for low-level vision tasks, while preserving the model’s strong semantic generation capabilities.

## 5 Shadow Removal

### 5.1 Background

Shadow removal aims to eliminate cast shadows from images and restore consistent illumination between shadow and non-shadow regions [47]. Shadows are caused by partial occlusion of light by scene objects and are ubiquitous in natural environments. Although shadows provide useful geometric cues for human perception, they often introduce strong intensity discontinuities, color distortions, and loss of texture details, which severely degrade the performance of downstream vision tasks such as object detection, tracking, and segmentation. Indeed, removing shadows from an image remains a fundamental yet highly challenging low-level vision task.

Early shadow removal methods [120] primarily relied on handcrafted priors and carefully designed statistical features, such as illumination consistency, gradient constraints, and region smoothness. These optimization-based approaches were built upon highly idealized physical and photometric assumptions, which rarely hold in real-world scenes with complex lighting, textured backgrounds, and soft shadow boundaries. Consequently, they often suffer from noticeable artifacts, especially around shadow boundaries, and fail to generalize to diverse real-world scenarios.

With the rapid development of deep learning, fully supervised CNN-based shadow removal methods have achieved remarkable progress by learning pixel-wise mappings from shadow images to their shadow-free counterparts using large-scale paired datasets. While these methods significantly improve visual quality on benchmark datasets, they heavily rely on expensive pixel-level annotations and often exhibit severe overfitting with limited generalization capability. More importantly, shadow removal is inherently a region-wise corrupted problem that involves strong contextual and structural priors.

The recent introduction of Nano Banana Pro demonstrates remarkable capabilities of generative models in visual tasks. In this context, we systematically evaluate the performance of Nano Banana Pro on the single-image shadow removal task. Through comprehensive comparisons with existing representative methods, we provide an in-depth analysis of its strengths and limitations in real-world applications.

## 5.2 Qualitative and Quantitative Results

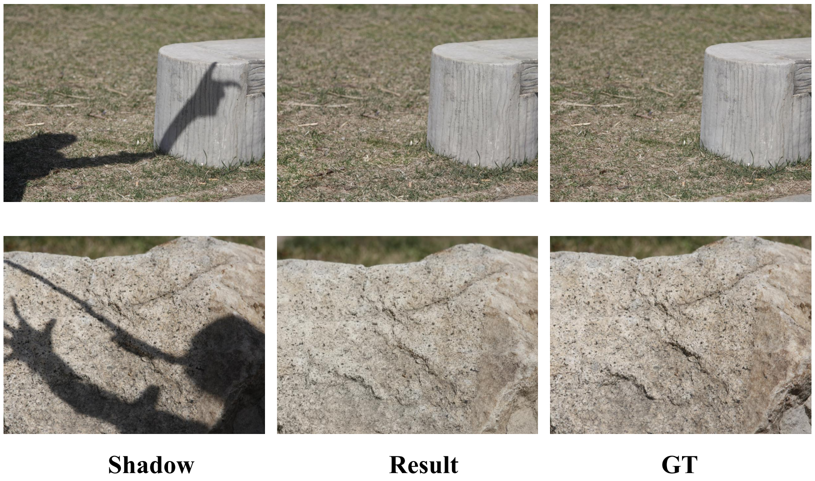


**Figure 13** Some well-performing visual examples of NB Pro on the SRD dataset [120] for the shadow removal task.

Fig. 13 presents the well-performing shadow removal results of NB Pro on the SRD dataset [120]. It can be observed that NB Pro effectively removes shadows from the image while highly preserving the original elements without alteration. Tab. 4 presents the quantitative comparison on SRD dataset of NB Pro against state-of-the-art shadow removal methods, using PSNR and SSIM as the primary evaluation metrics.

**Table 4** Quantitative comparisons on the SRD dataset [120]. The best results are highlighted by **black bold.**

| Method | PSNR ↑ | SSIM ↑ |
|---|---|---|
| DSC [59] (TPAMI'19) | 27.76 | 0.903 |
| DHAN [24](AAAI'20) | 30.51 | 0.949 |
| BMNet [211](CVPR'22) | 31.69 | 0.956 |
| ShadowFormer[45] (AAAI'23) | 32.90 | 0.958 |
| ShadowDiffusion [46] (CVPR'23) | 34.73 | 0.970 |
| HomoFormer [168] (CVPR'24) | **35.37** | **0.972** |
| Nano Banana Pro | 20.67 | 0.682 |

As shown in Tab. 4, a significant discrepancy exists between the visual quality discussed earlier and the numerical fidelity scores. While leading methods such as ShadowDiffusion [46] and HomoFormer [168] achieve PSNR scores exceeding 34 dB and SSIM values above 0.97, NB Pro records comparatively lower scores, with a PSNR of 20.67 dB and SSIM of 0.682. This quantitative gap can be attributed to the inherent characteristics of generative models: NB Pro prioritizes perceptual plausibility and visual naturalness over strict pixel-wise alignment with ground truth references. Unlike traditional methods that focus on precise reconstruction, generative approaches like NB Pro tend to produce images with enhanced visual appeal, which may deviate from the exact pixel values of ground truth images, resulting in lower scores on fidelity-based metrics.

In addition to its successful cases, we systematically document representative failure modes of NB Pro in shadow removal, as compiled in Fig. 14. These examples reveal two core, recurring limitations that stem from the model's generative nature: a propensity for over-generation and a blindness to subtle shadows.

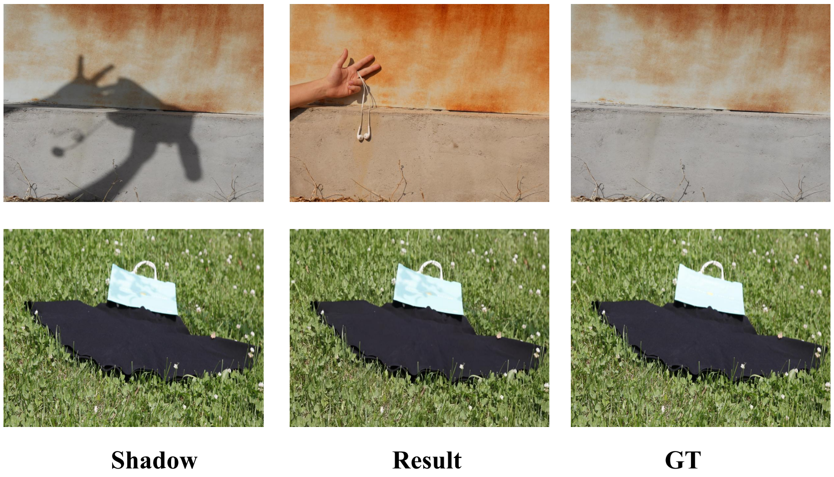


**Figure 14** Some visual failure examples of NB Pro on the SRD dataset [120] for the shadow removal task.

A consistent pattern of failure emerges across the SDR dataset, revealing critical limitations in NB Pro's approach. *The first failure mode stems from the model's inherent generative bias.* Driven by a compulsion to produce visually 'complete' scenes, NB Pro often prioritizes hallucinated content over fidelity. As illustrated in Fig. 14, while the cast shadow is successfully removed, the model erroneously synthesizes a new hand to fill the void, fundamentally compromising the scene's semantic integrity. This highlights a tension between the model's creative instinct and the strict fidelity required for low-level vision, where structural preservation is paramount. *The second failure mode exposes a lack of sensitivity in shadow detection.* NB Pro frequently overlooks soft, low-contrast, or faint shadows (as seen in the second example), leaving them entirely untreated. This suggests a potential bias in the training data or optimization objectives that under-weights subtle illumination changes. Furthermore, the model struggles with color fidelity, frequently exhibiting shifts in tone and saturation that deviate from the ground truth. Collectively, these structural hallucinations, detection failures, and color shifts lead to unsatisfactory quantitative performance.

## 5.3 Analysis

Nano Banana Pro demonstrates a remarkable ability to decouple shadow components from the underlying reflectance. However, its effectiveness is fundamentally constrained by its generative nature, which presents a significant paradigm mismatch with the strict fidelity requirements of shadow removal.

First, the model's inherent generative bias prioritizes perceptual plausibility over structural fidelity. As a generative model, NB Pro tends to hallucinate details, alter textures, or even synthesize new objects to create visually complete scenes. While this enhances aesthetic appeal, it compromises pixel-level accuracy, causing the output to deviate from a truthful reconstruction of the original reflectance. Consequently, NB Pro is better suited for creative applications requiring visual realism rather than low-level vision tasks demanding precise photometric or geometric correspondence.

Second, the model exhibits limited sensitivity in shadow detection. This limitation is likely attributable to biases in its training data or optimization objectives. If the training distribution under-represents subtle, soft, or low-contrast shadows, the model fails to learn the necessary features to identify them. As a result, NB Pro often leaves faint shadows untreated or, conversely, erroneously alters well-lit areas in an attempt to enforce uniform illumination, introducing artifacts in non-shadow regions.

Finally, these limitations are exacerbated by the incompatibility between generative outputs and traditional evaluation metrics. NB Pro typically generates images at fixed, high resolutions (e.g., 1K or 4K). Downsampling these outputs to match benchmark ground-truth resolutions smooths out high-frequency details, artificially depressing scores on pixel-wise metrics like PSNR and SSIM. More critically, a paradox arises where the model's generative enhancements, such as implicit super-resolution or denoising, are heavily penalized as deviations from the ground truth. This stark divergence between perceptual quality and quantitative scores underscores the inadequacy of current fidelity-based frameworks for evaluating generative restoration models.

# 6 Motion Deblurring

## 6.1 Background

Motion blur, caused by camera shake or fast-moving objects, remains one of the most pervasive artifacts degrading image quality in photography and computer vision tasks. Over the past decade, deep learning approaches have revolutionized dynamic scene deblurring. Early CNNs, such as DeepDeblur [115] and DeblurGAN [70], paved the way for more sophisticated architectures. Recently, Transformer-based models like Uformer [157] and Restormer [184] have dominated the field, achieving record-breaking scores in peak PSNR by effectively modeling long-range dependencies. However, these regression-based methods, typically optimized via MSE loss, tend to produce overly smooth results, often sacrificing high-frequency textures in favor of minimizing pixel-level error.

To overcome the "smoothing effect" and restore realistic details, the focus has shifted toward generative models, including GANs and Diffusion Models (e.g., ID-CDM [156], HI-Diff [20]). These approaches leverage strong generative priors, creating plausible textures for missing details. While they significantly enhance perceptual quality, they introduce a critical challenge: the perception-distortion trade-off. As the model strives to generate sharper and more visually pleasing images, it risks drifting away from the ground truth fidelity, potentially creating artifacts or "hallucinating" content that does not exist in the original scene.

In this technical report, we present a comprehensive evaluation of Nano Banana Pro (NB Pro). Our investigation reveals a fundamental limitation in this aggressive generative approach. While NB Pro demonstrates exceptional capability in synthesizing sharp textures for static environments and text, specifically in challenging low-light scenarios, it suffers from severe semantic instability. Our quantitative analysis (Tab. 5) and visual inspection confirm that the model's pursuit of visual sharpness often comes at the cost of fidelity to the input signal.

Specifically, we observe that NB Pro struggles with complex motion trajectory reduction, often misinterpreting motion cues as structural elements, which leads to ghosting artifacts. Furthermore, the model exhibits a tendency to alter semantic information, such as facial identities and text characters. These generative behaviors result in comparatively low quantitative scores (PSNR/SSIM). By analyzing NB Pro's performance across synthetic (GoPro [115], HIDE [132]) and real-world (RealBlur [127]) benchmarks, this report aims to dissect the failure modes of generative deblurring, highlighting the significant gap between producing visually plausible images and maintaining structural accuracy.

## 6.2 Qualitative Results

We conduct a visual analysis of NB Pro on standard benchmark datasets(GoPro [115], HIDE [132] and RealBlur [127]) to evaluate its performance in restoring structural details and handling complex degradations.

### 6.2.1 Performance on Synthetic Datasets

NB Pro demonstrates impressive deblurring capabilities on synthetic datasets, particularly in recovering static environmental details. As observed in Fig. 15, in the second column of GoPro and the first column of HIDE, the model effectively suppresses severe motion blur and restores high-frequency structures with great precision. Architectural elements, such as the building facades and pavement textures, are reconstructed with high fidelity. Notably, the model excels at text preservation in these scenarios; for instance, the "SEPHORA" signboard in the HIDE dataset is rendered clearly. This indicates strong spatial adaptability in handling rigid motion and structural edges.

However, significant limitations become apparent when processing highly dynamic scenes involving humans. The model struggles to fully eliminate complex synthetic motion trajectories, leading to noticeable residual artifacts. In the first column of GoPro, for example, the clothes hanging on the wall appear duplicated, and the woman's headscarf exhibits a double-layer ghosting effect, suggesting an incomplete resolution of the motion path. Furthermore, the model tends to hallucinate facial details. While the restored faces in both GoPro (1st column) and HIDE (2nd column) appear visually sharp, they suffer from semantic inconsistencies. The facial features are altered to the extent that the identity of the pedestrians no longer matches the Ground Truth (GT), highlighting a critical lack of fidelity in semantic reconstruction.

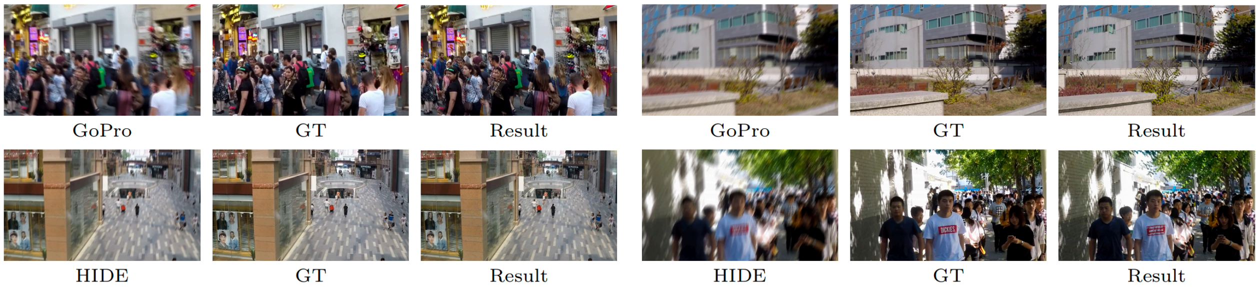


**Figure 15** Visual results of Nano Banana Pro on synthetic blur datasets (GoPro and HIDE).

### 6.2.2 Performance on Real-World Datasets

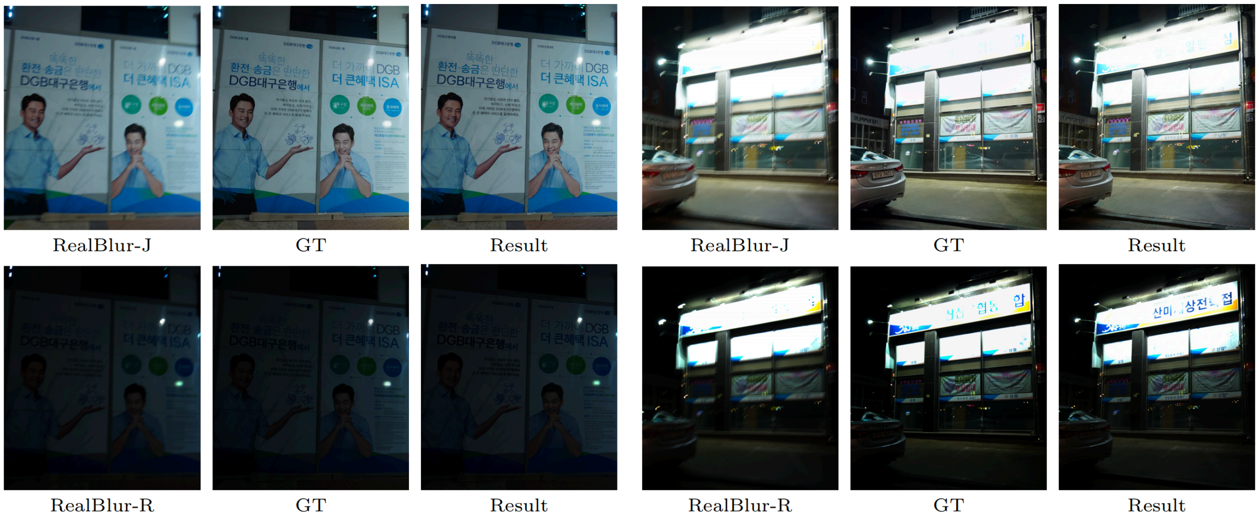
On real-world datasets, NB Pro exhibits strong robustness against complex degradations such as low light and overexposure. As seen in Fig. 16, in the RealBlur-J dataset, the model successfully recovers the legibility of text on posters and signboards. Its ability to handle high dynamic range scenes is particularly noteworthy, in the storefront examples (2nd columns of both RealBlur-J and RealBlur-R), the model manages high-contrast lighting effectively. However, the result generated by NB Pro in the second column of RealBlur-R deviates significantly from the GT properties. While the output appears cleaner, it aggressively removes noise and alters lighting textures, resulting in a synthesized appearance that loses the atmosphere of the original scene.

Moreover, the model's reliance on generative priors introduces substantial perceptual deviations from the ground truth. In the poster examples of RealBlur-J (1st column), the restored facial features differ from the original image, creating a "hallucinated" face that does not preserve the subject's identity. Similar discrepancies are observed in the text content of the RealBlur-J (2nd column) result, where the generated characters deviate from the GT, such as pink characters on the window. This can be also observed in the second column of RealBlur-R, where the characters on the illuminated sign are significantly altered compared to GT. Additionally, color fidelity is occasionally compromised. For instance, in the first column of RealBlur-R, the skin tone of the person in the result exhibits a noticeable color shift compared to the target image. These issues indicate that while NB Pro excels at producing visually pleasing results, it sacrifices faithfulness to the original semantic content.

## 6.3 Quantitative Results

Tab. 5 presents the quantitative comparison of NB Pro against state-of-the-art deblurring methods, including transformer-based models like Uformer and Restormer, as well as recent diffusion-based approaches like HI-Diff and ID-CDM. The evaluation is performed on four standard benchmarks: GoPro [115], HIDEHIDE [132] and RealBlur [127], using PSNR and SSIM as the primary metrics.

As observed in Tab. 5, a significant divergence exists between the previously discussed visual sharpness and the numerical fidelity scores. While top-performing methods such as ID-CDM and HI-Diff achieve PSNR scores exceeding 33 dB on the GoPro dataset and 36 dB on RealBlur-R, NB Pro records comparatively lower values, such as 21.41 dB on GoPro and 21.35 dB on HIDE. Similarly, the SSIM scores for NB Pro range between 0.645 and 0.778, whereas competing methods consistently score above 0.90. This quantitative gap can be primarily attributed to the model's heavy reliance on strong generative priors, which prioritizes perceptual plausibility over strict pixel-wise alignment with the ground truth.


RealBlur-J GT Result RealBlur-J GT Result

RealBlur-R GT Result RealBlur-R GT Result

**Figure 16** Visual results of Nano Banana Pro on real-world blur dataset (RealBlur-J and RealBlur-R).

The lower PSNR and SSIM scores directly corroborate the limitations identified in the qualitative analysis. Standard metrics like PSNR are highly sensitive to pixel-level deviations. As noted in the visual evaluation, NB Pro tends to hallucinate high-frequency details, such as altering facial identities or modifying text characters, to maximize sharpness. These generated features, while appearing visually coherent, act as "errors" regarding the reference image, leading to heavy penalties in signal-to-noise calculations. Furthermore, the reported semantic inconsistencies, such as color shifts and the "double-layer ghosting" effects caused by misinterpreted motion trajectories, significantly disrupt structural similarity, resulting in the observed drop in SSIM. This confirms that the model's generated content often diverges from the underlying ground truth signal.

**Table 5** Quantitative comparison results of Nano Banana Pro and other representative methods on four benchmarks.

| Method | GoPro | | RealBlur-R | | RealBlur-J | | HIDE | |
|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| DeepDeblur [115] | 29.08 | 0.914 | 32.51 | 0.841 | 27.87 | 0.827 | 25.73 | 0.874 |
| GAMD [102] | 33.14 | 0.9284 | 34.00 | 0.9265 | – | – | – | – |
| DeblurGAN [70] | 28.70 | 0.858 | 33.79 | 0.903 | 27.97 | 0.834 | 24.51 | 0.871 |
| DeblurGAN-v2 [71] | 29.55 | 0.934 | 35.26 | 0.944 | 28.70 | 0.866 | 26.61 | 0.875 |
| DBGAN [194] | 31.10 | 0.942 | 33.78 | 0.909 | 24.93 | 0.745 | 28.94 | 0.915 |
| Uformer-B [157] | 32.97 | 0.967 | 36.22 | 0.957 | 29.06 | 0.884 | 30.83 | **0.952** |
| Stripformer [140] | 33.08 | 0.962 | 36.08 | 0.954 | 28.82 | 0.876 | 31.03 | 0.940 |
| Restormer [184] | 32.92 | 0.961 | 36.19 | 0.957 | 28.96 | 0.879 | 31.22 | 0.942 |
| IR-SDE [104] | 30.70 | 0.901 | 33.96 | 0.918 | 24.21 | 0.729 | – | – |
| DiffIR [165] | 33.20 | 0.963 | – | – | – | – | 31.55 | 0.947 |
| HI-Diff [20] | **33.33** | 0.964 | 36.28 | **0.958** | **29.15** | **0.890** | 31.46 | 0.945 |
| ID-CDM [156] | 33.19 | **0.970** | **36.34** | 0.955 | 28.96 | 0.887 | **31.53** | 0.950 |
| **Nano Banana Pro** | **21.41** | **0.645** | **27.43** | **0.778** | **24.51** | **0.747** | **21.35** | **0.662** |

## 6.4 Analysis

The discrepancy between NB Pro's superior visual sharpness and its lower quantitative scores stems primarily from the perception-distortion trade-off. While regression-based methods minimize pixel-wise error to maximize PSNR, often resulting in over-smoothed textures, NB Pro leverages generative priors to create high-frequency details. This strategy produces visually realistic textures but introduces stochastic pixel deviations from the ground truth. Since standard metrics like PSNR penalize any deviation equally, NB Pro receives lower scores despite operating at a higher level of perceptual quality.

Furthermore, the evaluation on real-world datasets reveals the specific limitations of the model's reconstruction capability. While the ground truth in RealBlur sometimes contains noise or light scattering, NB Pro's tendency to completely "clean" these elements represents a deviation from the scene's authentic characteristics. Rather than simply restoring the signal, the model synthesizes a new, idealized version of the image. This leads to low quantitative scores not just because of metric limitations, but because the model fails to preserve the original distribution of the input data, effectively altering the scene's atmosphere.

Consequently, this reliance on generative priors introduces significant risks regarding semantic fidelity. When motion blur obliterates structural information, the model synthesizes plausible but factually incorrect content, leading to the observed "identity swaps" in human faces and ghosting artifacts in complex motion paths. While NB Pro excels in perceptual synthesis, making it suitable for visually-oriented restoration, these semantic inconsistencies highlight its unsuitability for applications requiring strict adherence to the original input signal, such as forensic analysis or high-fidelity surveillance.

# 7 Defocus Deblurring

## 7.1 Introduction

Defocus blur, an inherent optical phenomenon resulting from limited depth of field and aperture configurations, presents one of the most complex challenges in computational photography and low-level vision. Unlike uniform degradations such as global motion blur, defocus acts as a spatially varying aberration where the point spread function (PSF) changes according to the scene depth. This results in a non-uniform loss of high-frequency details and edge information, making the restoration process an ill-posed inverse problem that requires estimating spatially adaptive kernels.

The field of single-image defocus deblurring has advanced significantly with the advent of deep learning. Early data-driven approaches, such as DPDNet [2], established strong baselines by leveraging dual-pixel data to supervise defocus removal. Subsequent architectures, including IFANet [73] and KPAC [133], introduced iterative filtering and kernel prediction mechanisms to better handle spatially varying blur. More recently, the introduction of Transformer-based architectures, such as Restormer [184], and multi-stage networks like MPRNet [111], has pushed the boundaries of restoration fidelity by capturing long-range dependencies and global context. Current state-of-the-art methods, such as GGKMNet [124], further refine this process by integrating grouped kernel modeling to precisely invert the blurring process across complex depth maps.

In this section, we extend our evaluation of the Nano Banana Pro (NB Pro) to the domain of defocus deblurring. Unlike the aforementioned supervised methods, which are trained specifically on paired defocus datasets, our investigation focuses on assessing the NB Pro model in a zero-shot inference setting. By benchmarking the model against standard datasets such as DPDD [2] and RealDOF [129], we aim to analyze its efficacy in handling the inverse problem of deblurring without domain-specific fine-tuning. Specifically, we examine whether the model's processing pipeline can genuinely recover lost structural information comparable to established specialized networks, or if it merely relies on superficial enhancement techniques.

## 7.2 Quantitative Results

Tab. 6 presents the quantitative evaluation on the DPDD [2] and RealDOF [129] datasets. The results indicate a substantial performance gap between the Nano Banana Pro and established defocus deblurring methods. On the DPDD dataset, the Nano Banana Pro yields a PSNR of 20.180 dB and an SSIM of 0.635, significantly trailing the state-of-the-art GGKMNet by over 6 dB. A similar deficiency is evident on the RealDOF dataset, where NB Pro lags behind even early baselines like DPDNet. These low metrics align consistently with our qualitative findings: the depressed PSNR reflects the model's general failure to restore pixel-level sharpness, while the low SSIM corroborates the structural hallucinations and inability to remove blur observed in the visual comparison. However, it is worth noting that the competing methods, such as Restormer and DPDNet,

**Table 6** Quantitative comparison on DPDD and RealDOF datasets. The best results are highlighted in **black bold.**

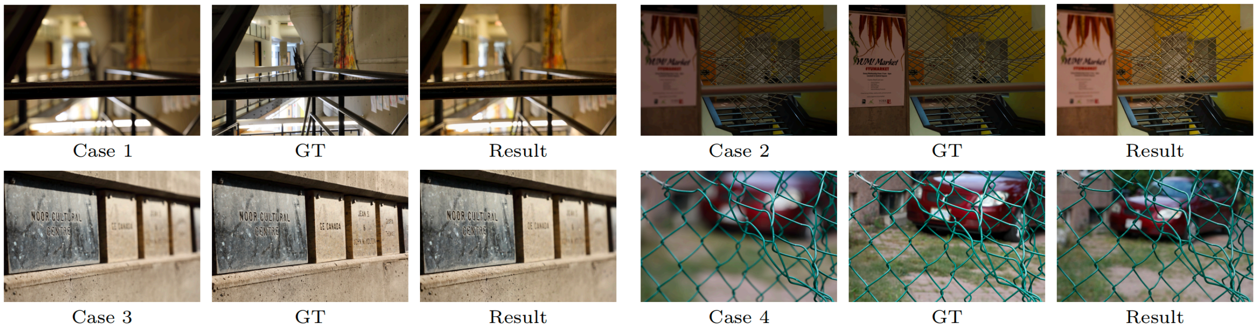
| Model | DPDD |  | RealDOF |  |
|---|---|---|---|---|
|  | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| DPDNet [2] | 24.348 | 0.747 | 22.870 | 0.670 |
| AIFNet [129] | 24.213 | 0.742 | 23.093 | 0.680 |
| IFANet [73] | 25.366 | 0.789 | 24.712 | 0.748 |
| KPAC [133] | 25.221 | 0.774 | 23.975 | 0.762 |
| GKMNet [121] | 25.468 | 0.789 | 24.254 | 0.732 |
| MDP [3] | 25.347 | 0.763 | 23.500 | 0.681 |
| DRBNet [130] | 25.485 | 0.792 | 24.884 | 0.751 |
| MPRNet [111] | 25.730 | 0.792 | 24.541 | 0.736 |
| Restormer [184] | 25.980 | **0.811** | 25.091 | 0.762 |
| INIKNet [123] | 26.055 | 0.803 | 25.231 | 0.765 |
| NRKNet [122] | 26.109 | 0.810 | 25.148 | 0.768 |
| GGKMNet [124] | **26.272** | 0.810 | **25.355** | **0.770** |
| **NB Pro** | **20.180** | **0.635** | **20.821** | **0.641** |

are supervised models trained directly on the DPDD dataset, whereas NB Pro is evaluated here in a zero-shot setting without domain-specific training.

## 7.3 Qualitative Results

We evaluated the perceptual performance of the Nano Banana Pro by conducting a comprehensive visual analysis on the DPDD and RealDOF datasets. The results indicate a consistent limitation in the model's ability to recover high-frequency details from severe defocus, with the model frequently prioritizing global contrast enhancement over effective blur removal.

On the DPDD dataset, NB Pro behaves more akin to an image enhancement filter than a specialized deblurring network. As seen in Fig. 17, in scenarios such as Case 1 and Case 2, the primary modification to the input is a global increase in luminance and contrast. While this improves the visual punch of the image, it fails to address the underlying degradation. Specifically, the severely defocused foreground in Case 2 remains blurry, and the background bokeh in Case 1 is only marginally reduced in spread. Furthermore, the model exhibits instability in structural reconstruction. This is evident in Case 3, where the restoration process hallucinates semantic details, incorrectly recovering the text "GE CANADA" as "OE CANADA" despite only a slight improvement in sharpness. Similarly, in Case 4, while the foreground fence is adequately sharpened, it compromises geometric fidelity, resulting in an unexplained scale alteration of the red vehicle in the background.



Case 1 GT Result Case 2 GT Result

Case 3 GT Result Case 4 GT Result

**Figure 17** Some representative qualitative results of Nano Banana Pro on the DPDD dataset.

The limitations of NB Pro are even more pronounced in the RealDOF dataset evaluations, where the model demonstrates a negligible deblurring effect across multiple test cases. As seen in Fig. 18, in scenes with spatially varying blur, such as the mid-range focus in Case 1 and the foreground focus in Case 2, the model fails to reverse the defocus entirely. The output images are characterized solely by a slight boost in contrast, leaving the blurred regions perceptually identical to the input. While the model achieves a degree of sharpness recovery in the fully blurred scenario of Case 3, this comes at the cost of introducing high-frequency artifacts, specifically salt-and-pepper noise visible on the building structures. Moreover, the restoration in Case 4 is depth-limited. The model successfully restores the ground texture closest to the lens but fails to extend the depth of field to the background, which remains in a state of defocus with only a minor reduction in the circle of confusion. Collectively, these qualitative results suggest that the Nano Banana Pro lacks the robustness required for consistent defocus deblurring, often failing to produce a discernible improvement in image sharpness.

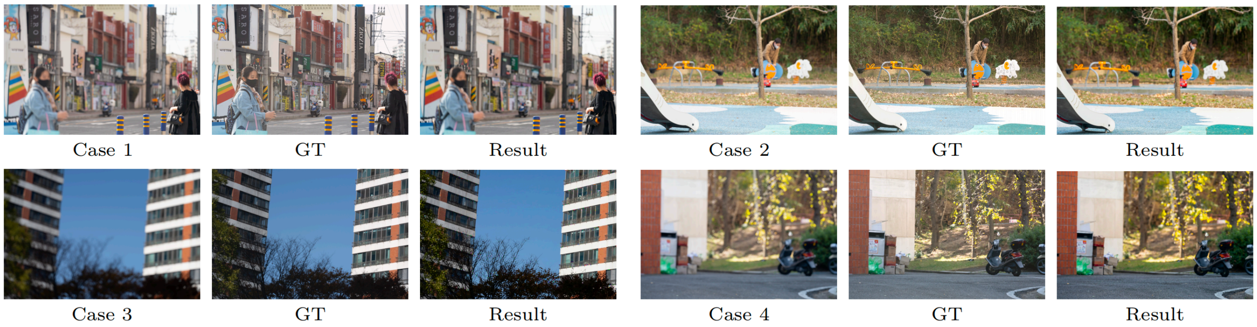


**Figure 18** Some representative qualitative results of Nano Banana Pro on the RealDOF dataset.

## 7.4 Analysis

The disparity between the quantitative metrics and the qualitative visual outputs reveals the inherent instability of applying a general-purpose generative model to the specific physical constraints of defocus deblurring. Our analysis suggests that the Nano Banana Pro does not perform a mathematical inversion of the optical point spread function. Instead, it relies on semantic-aware generative priors to synthesize sharp details. This results in a bimodal behavior where the model oscillates between superficial contrast adjustment and aggressive, perceptually driven reconstruction depending on the scene's semantic recognisability.

In complex scenes with varying depth and high-frequency clutter, such as those in the DPDD dataset, the model's generative mechanism often struggles to identify coherent structural cues. Consequently, the model defaults to a global contrast maximization approach. This explains the consistently low PSNR values observed in Tab. 6. While increasing local contrast can improve perceptual punch in slightly out-of-focus regions, it fails to mathematically invert the point spread function. Consequently, the model exhibits inconsistent restoration behaviors, resorting to superficial contrast adjustment when semantic cues are ambiguous, while attempting aggressive reconstruction in scenes with recognizable structures.

However, in scenarios with regular, recognizable structures such as the building facade in Fig. 18 Case 3, the model successfully engages its learned priors to "re-paint" the geometry, effectively removing the blur. Yet, this reconstruction is perceptually driven rather than physically constrained, leading to high-frequency artifacts. The salt-and-pepper noise observed along the window frames is likely a byproduct of the generative process (e.g., instability in the diffusion sampling) attempting to force high-frequency gradients into latent features that do not perfectly align with the degraded input.

Conversely, when the defocus aligns with typical photographic aesthetics, the model exhibits passivity. This is clearly observed in Fig. 18 Case 4, where the background exhibits only mild defocus and remains semantically distinguishable, yet the model sharpens only the foreground ground texture while leaving the background blur intact. This divergence strongly suggests that the model's priors interpret the slight background defocus as an intended aesthetic attribute, specifically, as a natural depth-of-field, rather than a degradation requiring correction. Unlike dedicated deblurring networks that aim to minimize the circle of confusion globally, NB Pro appears to prioritize perceptual naturalness, effectively treating the background blur as context to be preserved rather than an error to be inverted.

Finally, the lack of fidelity constraints in this zero-shot setting leads to significant structural deviations. Since the model prioritizes perceptual plausibility over pixel-wise accuracy, it introduces semantic errors when input ambiguity is high (such as hallucinating "OE CANADA" instead of "GE CANADA" in Fig. 18 and altering the geometric scale of the vehicle in Fig. 17). These behaviors confirm that the Nano Banana Pro operates as an image re-synthesis engine rather than a dedicated restoration tool, resulting in low quantitative scores (PSNR/SSIM) despite occasional visual successes.

# 8 Denoising

## 8.1 Background

Recent advancements in vision-language models [7, 94, 100] mark a significant paradigm shift toward unified architectures capable of integrating diverse modalities and tasks within a single framework. Notably, models such as Nano Banana Pro have demonstrated that the synergistic training of understanding and generation objectives can unlock emergent capabilities and enhance cross-task generalization.

Despite the architectural elegance and representation efficiency offered by unified models [100], a critical question persists: Can these generalist systems rival the precision of dedicated restoration networks [114] in specialized low-level vision tasks? Image denoising, specifically, stands as a rigorous litmus test. It evaluates a model's capacity to preserve fine-grained details, textures, and structural fidelity, which are intrinsic not only to restoration but also to high-fidelity image generation and editing.

In this technical report, we conduct a systematic evaluation of Nano Banana Pro's denoising performance across five established benchmark suites: McMaster [195] for natural image statistics, Kodak24 [34] for photographic quality assessment, Urban100 [61] for challenging high-frequency texture reconstruction, and PolyU [171] and SIDD-small [1] for real-world sensor noise suppression. This study serves a dual purpose: first, to determine whether Nano Banana Pro's unified training regime yields competitive low-level restoration quality; and second, to elucidate the interplay between generative capabilities and fine-grained reconstruction, providing insights to guide the design of future unified architectures.

## 8.2 Experimental Setup

NanoBanana is a closed-source unified multimodal model accessed through its official API. As a model capable of image understanding, generation, and text-driven editing, we evaluate its denoising capability by providing noisy images alongside a natural language instruction. The prompt used throughout our experiments is: "This is a noisy image, please remove the noise in this image while keep other elements in this image unchanged."

**Datasets.** We evaluate Nano Banana Pro on five widely-used image denoising benchmarks spanning both synthetic and real-world noise scenarios. For synthetic noise datasets like McMaster [195], Kodak24 [34], and Urban100 [61], we corrupt clean images with additive white Gaussian noise at a fixed noise level of $\sigma = 50$, representing a challenging high-noise regime. McMaster contains 18 high-resolution images with rich color and texture, Kodak24 comprises 24 classic photographic images, and Urban100 includes 100 images with complex urban structures and repetitive patterns that stress high-frequency reconstruction. For real-world noise datasets like PolyU [171] and SIDD-small [1], we use the standard noisy/clean image pairs provided by each benchmark without additional synthetic corruption. These datasets capture realistic sensor noise from various camera devices under diverse lighting conditions, presenting a more practical evaluation scenario.

**Resolution and Failure Case Handling.** Nano Banana Pro outputs images at approximately 1K resolution, though the exact dimensions vary across samples (e.g., 1024×1024, 1200×896, 720×1456). We resize the output images to match the resolution of corresponding ground truth images using bilinear interpolation. All metrics are then computed between the resized outputs and the ground truths. In addition, during evaluation, we observed that Nano Banana Pro occasionally produces outputs that are either semantically irrelevant to the input image or fail to remove noise effectively. In such cases, we regenerate the output by resubmitting the same input and prompt to the API until a valid denoised result is obtained. This protocol ensures that our quantitative metrics reflect the model's denoising capability under successful generation, while the occurrence of such failures is noted as a limitation of applying unified generative models to restoration tasks.

**Evaluation Metrics.** We adopt two complementary metrics to assess denoising quality: PSNR (Peak Signal-to-Noise Ratio), which measures pixel-level fidelity. SSIM (Structural Similarity Index): Evaluates perceptual structural similarity. Both are computed on RGB channels

## 8.3 Quantitative and Qualitative Results

To systematically evaluate the image denoising capabilities of Nano Banana Pro, we invoked the model via its official API and compared its performance against five representative task-specific baselines (DnCNN [192], Restormer [184], MaskDenoising [15], HAT [174], and DIL [89]). The evaluation spans two distinct regimes. First, we employed three synthetic benchmarks—McMaster [195], Kodak24 [34], and Urban100 [61]—corrupted with additive Gaussian noise ($\sigma = 50$) to test reconstruction across varying complexities. Specifically, McMaster assesses basic noise removal in smooth textures; Kodak24 covers diverse natural scenes to balance texture and color fidelity; and Urban100 challenges the model's ability to preserve high-frequency details within complex architectural structures. Complementing these synthetic tests, we assessed real-world blind denoising performance using SIDD Val [1] and PolyU [171], where no prior noise information is provided. SIDD Val serves as a core benchmark for handling authentic sensor noise captured under varying lighting and device conditions. Furthermore, PolyU is utilized to stress-test the model's generalization capabilities on irregular noise distributions characteristic of low-light and complex environments.

**Table 7** Quantitative results of performance comparison on synthetic and natural noise datasets. The metrics are PSNR and SSIM, where higher values indicate better performance.

| Noise Types | Datasets | DnCNN [192] | Restormer [184] | MaskDenoising [15] | HAT [174] | DIL [89] | NB pro |
|---|---|---|---|---|---|---|---|
| Gauss $\sigma = 50$ | McMaster [195] | 20.18/0.312 | 20.47/0.312 | 20.63/0.379 | 20.79/0.364 | 26.61/0.669 | 21.57/0.594 |
| | Kodak24 [34] | 19.78/0.301 | 20.12/0.321 | 20.72/0.368 | 21.04/0.390 | 27.46/0.736 | 20.04/0.517 |
| | Urban100 [61] | 19.62/0.420 | 19.36/0.437 | 20.51/0.485 | 20.80/0.492 | 25.89/0.768 | 19.22/0.607 |
| Natural | SIDD Val [1] | -/- | -/- | 33.14/0.913 | 28.58/0.570 | 34.76/0.848 | 26.76/0.681 |
| | PolyU [171] | -/- | -/- | 24.78/0.812 | 37.25/0.948 | 37.65/0.950 | 22.82/0.806 |

As shown in Tab. 7, Nano Banana Pro exhibits a substantial performance deficit compared to all task-specific baselines. On synthetic datasets, it lags significantly behind the state-of-the-art DIL, with PSNR gaps ranging from 5.04 dB to 7.42 dB and SSIM reductions between 0.075 and 0.219. Notably, this disparity persists regardless of texture complexity (from McMaster to Urban100), indicating a fundamental lack of competitiveness in Gaussian noise removal. This limitation is further exacerbated in real-world blind denoising tasks. On SIDD Val [1], Nano Banana Pro trails DIL by 8.00 dB in PSNR. The gap widens drastically on PolyU [171], where it underperforms DIL by 14.83 dB and even falls behind the basic MaskDenoising model. These results underscore an inherent inability of Nano Banana Pro to effectively model and remove complex, realistic noise compared to specialized restoration models.

Fig. 19 visually compares Nano Banana Pro against state-of-the-art baselines. The results reveal a distinct characteristic of the generative approach: a trade-off between perceptual clarity and pixel-level fidelity. As shown in the first row, Nano Banana Pro exhibits exceptional perceptual quality on text-rich images. Leveraging its generative priors, it reconstructs the characters with remarkable sharpness. Notably, the output appears even clearer and more legible than the Ground Truth, effectively performing text enhancement alongside denoising. Conversely, the model struggles with consistency in texture and color, as seen in the subsequent rows: In the second row, the model fails to recover the subtle grain of the surface. Instead of preserving the original high-frequency details, it produces an over-smoothed or hallucinated texture that deviates significantly from the Ground Truth. In the third row, the model introduces chromatic deviations. While the noise is removed, the color of the grapes shifts noticeably (appearing brighter and yellower). These cases underscore that while Nano Banana Pro can generate visually pleasing results, it lacks the strict fidelity required for high-precision restoration tasks.

## 8.4 Discussion

Based on the experimental results and architectural characteristics, the suboptimal denoising performance of Nano Banana Pro is attributed to two primary factors:

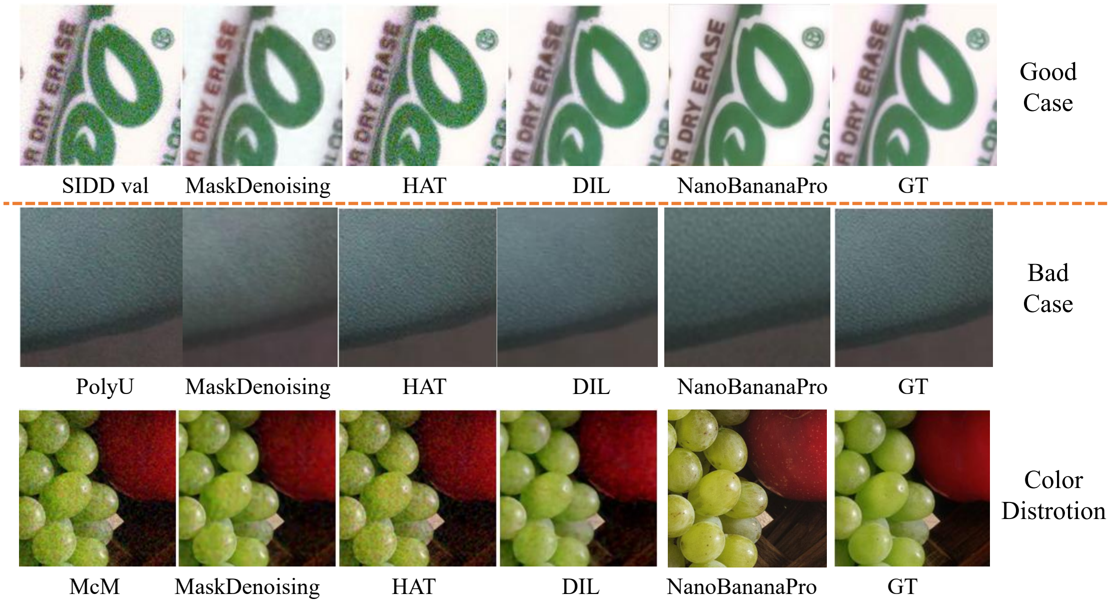


**Figure 19** Visual comparison of denoising results. The first row shows a successful case, where Nano Banana Pro produces text with sharper edges than the Ground Truth, partly due to its generative priors. The last two rows, however, reveal limitations in fidelity: in the second row, surface texture is not preserved, losing high-frequency details; in the third row, noticeable color distortion shifts the hue of the grapes away from the Ground Truth.

**Misalignment of Task Objectives**: Nano Banana Pro is a general-purpose model optimized for high-level multimodal understanding and generation, rather than low-level pixel-wise restoration. It lacks the specialized architectural biases and targeted loss functions that enable baseline models to effectively balance noise removal with detail preservation.

**Trade-off Between Generative Prior and Pixel Fidelity**: As a unified model, Nano Banana Pro prioritizes semantic plausibility and visual coherence over strict pixel-level accuracy. This generative nature often leads to the over-smoothing of high-frequency details in pursuit of "reasonable" content, resulting in inferior quantitative metrics compared to task-specific models trained via strict supervision.

In summary, this study evaluated the unified generative model Nano Banana Pro against state-of-the-art specialized models on both synthetic Gaussian noise and real-world blind denoising datasets. Nano Banana Pro significantly underperforms task-specific baselines across all benchmarks, indicating limited competitiveness in direct denoising applications. Direct application of Nano Banana Pro for denoising is not recommended without modification. Enhancing its utility requires targeted adaptations such as prompt engineering, parameter fine-tuning, or integration with post-processing modules. Future research should explore methodologies to align general-purpose generative priors with low-level processing demands.

# 9 Reflection Removal

## 9.1 Background

In fields such as computer vision, clear and interference-free image data is a fundamental foundation for subsequent analytical tasks including object detection and semantic segmentation. However, reflective surfaces like glass, water, and metal easily reflect ambient light into images, forming an interfering reflection layer over the real scene. This causes blurred details and obscured target information, directly compromising the reliability and accuracy of subsequent tasks. Single-Image Reflection Removal (SIRR), as a core technical solution, aims to accurately separate the transmission layer (real scene) from the reflection layer (interfering component) in a single mixed image to restore the true scene. It holds irreplaceable practical value in autonomous driving, security monitoring, consumer electronics, and other areas.

SIRR is inherently a typical ill-posed inverse problem—without additional constraints, mixed image decomposition has infinitely many solutions. Early traditional methods relied on manually designed prior knowledge (sparsity, smoothness, and other assumptions) and linear modeling (such as $I = T + R$) to simplify the problem, but real-world reflections exhibit complex nonlinear characteristics due to light intensity, shooting angle, surface material, and other factors, leading to limited generalization of these methods. In recent years, deep learning has become the mainstream in SIRR research, forming three core architectures: single-stage approaches that directly output the target layer via a single network [161, 198], two-stage approaches that perform intermediate feature estimation followed by refinement [30, 91], and multi-stage approaches that achieve reflection removal through recurrent cascaded iterative optimization [77, 176]. Notably, the rapid development of generative artificial intelligence has injected new vitality into the field, with methods based on diffusion models and Transformers demonstrating potential to break through traditional limitations [52, 182]. Nevertheless, existing approaches face significant bottlenecks: the scarcity of high-quality annotated datasets restricts model generalization, and issues such as scene information loss from strong reflections and overlapping appearance distributions between transmission and reflection layers make it hard to balance thorough reflection removal and detail preservation.

Existing research covers traditional methods, deep learning architectures, and generative paradigms, but there remains substantial room for improvement in robustness and detail fidelity under complex real-world scenarios. On one hand, while generative models show promise, their hallucination suppression capabilities and adaptability to complex reflection mechanisms in high-fidelity tasks like SIRR have not been fully verified. On the other hand, efficient solutions for diverse reflection scenarios (diverse material surfaces, extreme lighting, and other scenarios) are lacking, demanding more generalizable generative models. Based on this, this report focuses on the latest generative model Nano Banana Pro. By systematically comparing it with existing baselines using quantitative and qualitative metrics, we investigate its detail restoration, anti-interference performance, and generalization in real reflection removal scenarios. The goal is to reveal its core advantages and limitations, providing practical references for technical optimization and model design in the SIRR field.

## 9.2 Quantitative Results

To evaluate the performance of Google's Nano Banana Pro model on the SIRR task, we conducted experiments using the model as an off-the-shelf solution via API calls. Given the closed-source nature of the model which precludes task-specific fine-tuning, we adopted a direct inference strategy. Only the raw reflection-contaminated images and task-specific prompts were provided as input, without introducing any additional priors or auxiliary guidance. It is worth noting that the resolution of images generated by Nano Banana Pro is fixed at a scale of approximately 1024 pixels, which differs from the original input dimensions. To ensure the fairness and accuracy of the quantitative evaluation, all output images were resized to match the original resolution of the corresponding ground-truth images before metric calculation.

For a comprehensive assessment, we adopted three mainstream datasets in the SIRR domain as our evaluation benchmark. Specifically, we utilized Real20 [198], which contains 20 images from real-world glass reflection scenes; Nature [77], consisting of 20 samples focusing on outdoor natural landscapes; and SIR$^2$ [143], where we evaluated on its Objects, Postcard, and Wild subsets. We compared Nano Banana Pro against 15 state-of-the-art baseline models, including ERRNet [161], IBCLN [77], YTMT [56], Dong et al. [30], DSRNet [57], RAGNet [91], RRW [212], DSIT [58], RDNet [203], F2T2-HiT [13], Huang et al. [63], L-DiffER [52], DAI [55], Lu et al. [101], and WindowSeat [182].

The evaluation metrics assess both basic image quality and perceptual quality. For pixel-level fidelity, we employed Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), with comprehensive comparison results summarized in Tab. 8. To better assess visual perception, we further incorporated Multi-Scale Structural Similarity (MS-SSIM) and Learned Perceptual Image Patch Similarity (LPIPS). Note that lower LPIPS values indicate better perceptual quality. For these perceptual metrics, we selected recent SOTA methods for comparison (with baseline data sourced from WindowSeat [182]), as presented in Tab. 9. All metrics were calculated on the RGB channels between the resized output and the ground truth.

As shown in Tab. 8, Nano Banana Pro exhibits a notable performance gap compared to state-of-the-art specialist methods. Quantitatively, it lags behind across all datasets in pixel-wise metrics (PSNR/SSIM). This disparity largely stems from the fundamental difference in optimization objectives: regression-based SOTA

**Table 8** Quantitative Comparison of Single-Image Reflection Removal Methods. The best and second-best results are highlighted by **black bold** and <u>underline</u>, respectively. †: Training data includes Nature dataset; ⋆: Generative AI-based method. Note: ↑ indicates higher is better. SIR$^2$ dataset is divided into Objects, Postcard and Wild subsets; SIR$^2$(454) denotes the commonly used subset, while SIR$^2$(500) denotes the full public dataset.

| Method | Year | Real20 (20) | | Nature (20) | | Objects (200) | | Postcard (199) | | Wild (55) | | SIR$^2$ (454) | | SIR$^2$ (500) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| ERRNet [161] | 2019 | 22.89 | 0.803 | - | - | 24.87 | 0.896 | 22.04 | 0.876 | 24.25 | 0.853 | 23.55 | 0.882 | - | - |
| IBCLN [77] | 2020 | 21.86 | 0.762 | 23.57 | 0.783 | 24.87 | 0.893 | 23.39 | 0.875 | 24.71 | 0.886 | 24.20 | 0.884 | - | - |
| YTMT [56] | 2021 | 23.26 | 0.806 | - | - | 24.87 | 0.896 | 22.91 | 0.884 | 25.48 | 0.890 | 24.08 | 0.890 | - | - |
| Dong et al.† [30] | 2021 | 23.34 | 0.812 | 23.45 | 0.808 | 24.36 | 0.898 | 23.72 | 0.903 | 25.73 | 0.902 | 24.25 | 0.901 | - | - |
| DSRNet (w/o extra) [57] | 2023 | 24.23 | 0.820 | - | - | 26.28 | 0.914 | 24.56 | 0.908 | 25.68 | 0.896 | 25.45 | 0.909 | - | - |
| DSRNet (with extra) [57] | 2023 | 23.91 | 0.818 | - | - | 26.74 | 0.920 | 24.83 | 0.911 | 26.11 | 0.906 | 25.83 | 0.914 | - | - |
| RAGNet [91] | 2023 | 22.95 | 0.793 | - | - | 26.15 | 0.903 | 23.67 | 0.879 | 25.52 | 0.880 | 24.99 | 0.890 | - | - |
| RRW [212] | 2024 | 23.82 | 0.817 | 25.96 | 0.843 | - | - | - | - | - | - | 25.45 | 0.910 | - | - |
| DSIT (data I) [58] | 2024 | 25.06 | 0.836 | - | - | 26.81 | 0.919 | 25.63 | 0.924 | 27.06 | 0.910 | 26.32 | 0.920 | - | - |
| DSIT (data II) [58] | 2024 | 25.22 | 0.836 | - | - | 27.27 | **0.932** | 25.58 | 0.922 | 27.40 | 0.918 | 26.54 | 0.926 | - | - |
| RDNet (w/o nature) [203] | 2025 | 24.43 | 0.835 | - | - | 25.76 | 0.905 | 25.95 | 0.920 | 27.20 | 0.910 | 26.02 | 0.912 | - | - |
| RDNet (w nature) [203] | 2025 | 25.58 | 0.846 | - | - | 26.78 | 0.921 | 26.33 | 0.922 | 27.70 | 0.915 | 26.69 | 0.921 | - | - |
| F2T2-HiT [13] | 2025 | 21.64 | 0.766 | 26.08 | 0.837 | - | - | - | - | - | - | 25.72 | 0.903 | - | - |
| Huang et al. [63] | 2025 | 25.12 | 0.828 | 27.03 | 0.853 | 27.07 | 0.930 | 26.43 | 0.931 | 27.96 | **0.922** | 26.90 | 0.929 | - | - |
| L-DiffER⋆ [52] | 2025 | 23.77 | 0.821 | 23.95 | 0.831 | - | - | - | - | - | - | - | - | 25.18 | 0.911 |
| DAI⋆ [55] | 2025 | 25.24 | 0.840 | 27.05 | 0.846 | - | - | - | - | - | - | - | - | 27.32 | 0.931 |
| Lu et al.⋆ [101] | 2025 | - | - | - | - | - | - | - | - | - | - | - | - | 28.41 | 0.912 |
| WindowSeat⋆ [182] | 2025 | 26.28 | 0.856 | 27.12 | 0.849 | 28.81 | 0.944 | 29.17 | 0.934 | 28.97 | 0.935 | 28.99 | 0.939 | 28.75 | **0.940** |
| WindowSeat⋆ (Qwen-IE) [182] | 2025 | **26.60** | **0.864** | **27.57** | **0.855** | **28.85** | <u>0.938</u> | <u>28.70</u> | <u>0.933</u> | **29.44** | **0.936** | <u>28.84</u> | <u>0.936</u> | <u>28.60</u> | <u>0.937</u> |
| **Nano Banana Pro** | **2025** | **20.26** | **0.655** | **21.48** | **0.723** | **21.95** | **0.751** | **19.29** | **0.675** | **23.58** | **0.798** | **20.98** | **0.724** | **21.11** | **0.730** |

methods are supervised to minimize pixel-level reconstruction error, ensuring precise alignment. In contrast, the generative approach of Nano Banana Pro prioritizes semantic coherence over structural fidelity, often resulting in global intensity scaling and spatial shifts that heavily penalize PSNR, even if the image content is semantically correct.

**Table 9** Perceptual Quality Comparison (MS-SSIM and LPIPS) on Mainstream Datasets. ↑ indicates higher is better, while ↓ indicates lower is better. The best and second-best results are highlighted in **black bold** and <u>underline</u>. Baseline results are sourced from WindowSeat [182].

| Method | Real20 (20) | | Nature (20) | | Objects (200) | | Postcard (199) | | Wild (55) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MS-SSIM↑ | LPIPS↓ | MS-SSIM↑ | LPIPS↓ | MS-SSIM↑ | LPIPS↓ | MS-SSIM↑ | LPIPS↓ | MS-SSIM↑ | LPIPS↓ |
| DSRNet [57] | 0.8737 | 0.1831 | 0.9144 | 0.1478 | 0.9564 | 0.0847 | 0.9263 | 0.1260 | 0.9338 | 0.1096 |
| DAI [55] | 0.9045 | 0.1790 | 0.9309 | 0.2161 | 0.9638 | 0.0689 | 0.9567 | 0.1029 | 0.9423 | 0.0941 |
| RDNet [203] | 0.9081 | 0.1442 | 0.9231 | <u>0.1361</u> | 0.9609 | 0.0836 | 0.9361 | 0.1121 | 0.9406 | 0.0992 |
| DSIT [58] | 0.8934 | 0.1618 | 0.9223 | 0.1598 | 0.9586 | 0.0939 | 0.9441 | 0.1242 | 0.9447 | 0.0967 |
| WindowSeat [182] | <u>0.9296</u> | <u>0.1131</u> | <u>0.9435</u> | 0.1368 | **0.9759** | **0.0470** | **0.9693** | **0.0504** | <u>0.9625</u> | **0.0632** |
| WindowSeat (Qwen-IE) [182] | **0.9396** | **0.1074** | **0.9494** | **0.1355** | <u>0.9661</u> | <u>0.0550</u> | <u>0.9664</u> | <u>0.0549</u> | **0.9655** | <u>0.0682</u> |
| **Nano Banana Pro** | **0.8013** | **0.2411** | **0.8580** | **0.1851** | **0.8861** | **0.1552** | **0.8373** | **0.2513** | **0.8874** | **0.1578** |

Tab. 9 further illustrates the performance in terms of perceptual quality metrics. Despite the generative nature of Nano Banana Pro, which typically favors perceptual scores, it still exhibits high LPIPS values (e.g., 0.2513 on Postcard vs. 0.0549 for SOTA). Unlike PSNR, which penalizes misalignment, the poor LPIPS performance points to a deeper issue: *semantic and stylistic deviation*. The model tends to perform aggressive "image-to-image translation" rather than faithful restoration, altering fundamental scene characteristics—such as modifying illumination, hallucinating textures, or shifting the color domain—thereby drifting away from the ground truth's perceptual manifold.

From an interpretive perspective, the elevated LPIPS and sub-optimal MS-SSIM scores highlight the limitations of general-purpose generative models in high-fidelity restoration tasks. Although the model possesses strong generative capabilities, it lacks a precise mechanism for decoupling reflection layers from the background. The high LPIPS values suggest a substantial deviation in color distribution, texture details, and high-level semantic features relative to the ground truth. This deviation likely stems from the model partially merging residual reflections into the background or altering the original color and complex textures during the regeneration process. Consequently, while the generated images may appear visually natural, they fail to meet the strict fidelity requirements essential for reflection removal tasks.

### 9.3 Qualitative Results

In this section, we conduct a comprehensive qualitative evaluation of the proposed Nano Banana Pro. To establish a comparative baseline, we first benchmark our visual results against existing state-of-the-art methods in Fig. 20. Subsequently, we examine the specific restoration capabilities of our model through selected samples in Fig. 21. Finally, to ensure a balanced assessment and facilitate future improvements, we provide an analysis of the model's limitations by categorizing typical failure cases and degradation patterns in Fig. 22.

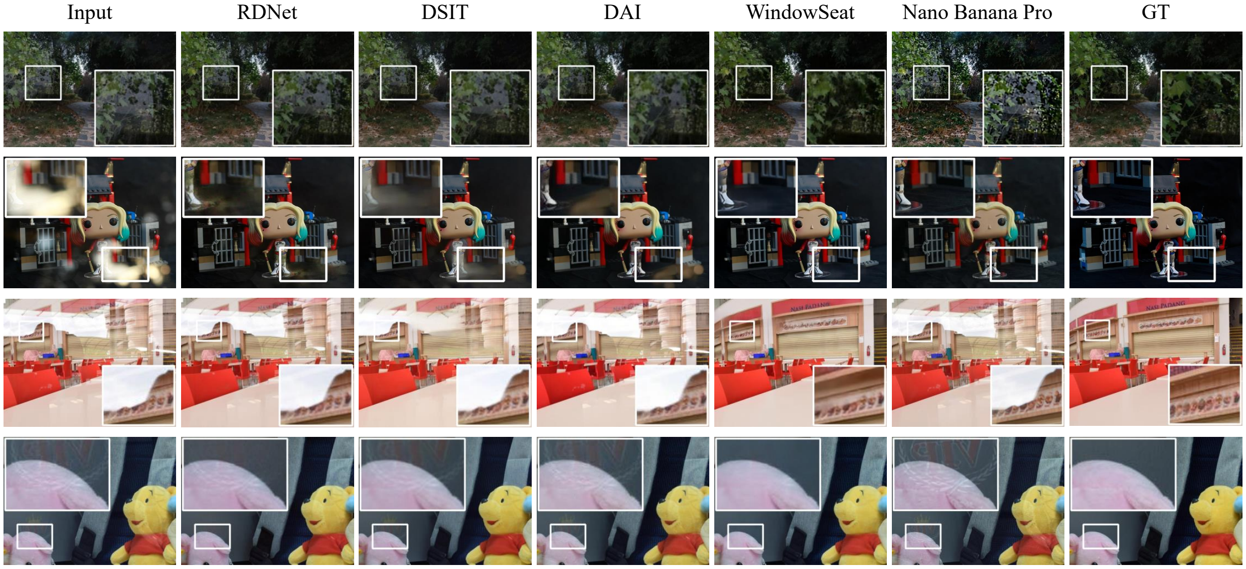


**Figure 20** Qualitative comparison of reflection removal results. Each row shows the input image, predictions from state-of-the-art methods, Nano Banana Pro, and the ground truth (GT) for a single sample. Except for the results of Nano Banana Pro, all other method results are sourced from WindowSeat [182].

As illustrated in Fig. 20, the proposed Nano Banana Pro exhibits a high variance in performance compared to state-of-the-art methods. In specific instances, our method outperforms existing approaches, yielding results that are visually comparable from the ground truth. However, generally, it lacks the stability of specialized regression-based models. Notably, the model struggles with preserving high-frequency details, leading to the loss of complex textures (e.g., row 1), or suffers from semantic ambiguity where reflection artifacts are erroneously interpreted as background elements and subsequently enhanced (e.g., row 4). These limitations contribute to a lower average performance despite the high perceptual quality in successful cases.

Fig. 21 showcases selected samples where Nano Banana Pro demonstrates superior restoration capabilities. It is observed that when there is a significant semantic or visual distinction between the reflection and transmission layers, the model effectively suppresses the reflection while preserving background integrity. The results indicate that the model possesses a high performance upper bound, occasionally achieving reconstruction quality nearly identical to the ground truth. We attribute this potential to the robust generative priors acquired from large-scale pre-training. However, the absence of domain-specific supervision for reflection separation implies a trade-off: without explicit guidance, the model may misapply these priors, failing to disentangle the layers or introducing generative hallucinations and noise into the transmission layer. Experimental results suggest that such misuse of priors accounts for a considerable portion of the suboptimal outputs.

Given the quantitative gap observed in the previous section, explicitly analyzing the failure modes provides critical insights into the misbehavior of generative priors. Based on the distinct characteristics of the introduced artifacts and degradation mechanisms, the suboptimal performance of Nano Banana Pro can be systematically categorized into six types, as visualized in Fig. 22.

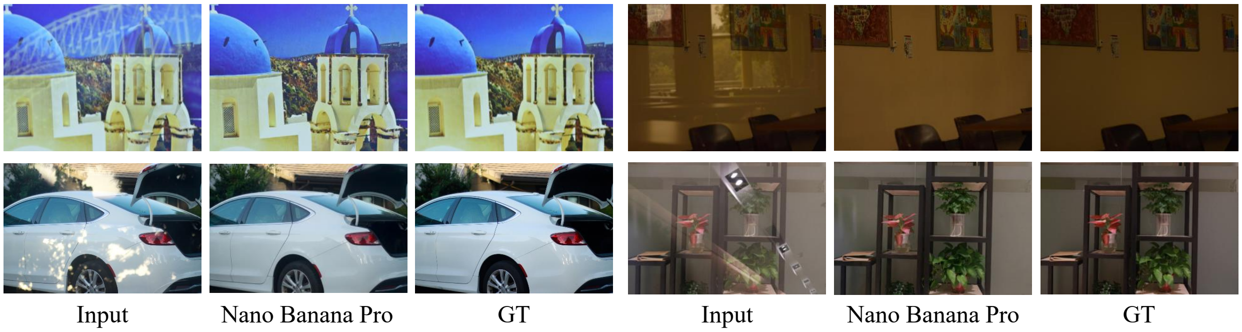


**Figure 21** Visual comparison of selected samples. We present the input images, the restoration results generated by Nano Banana Pro, and the corresponding ground truth. These examples illustrate the visual performance of the method in recovering background content across different scenes.

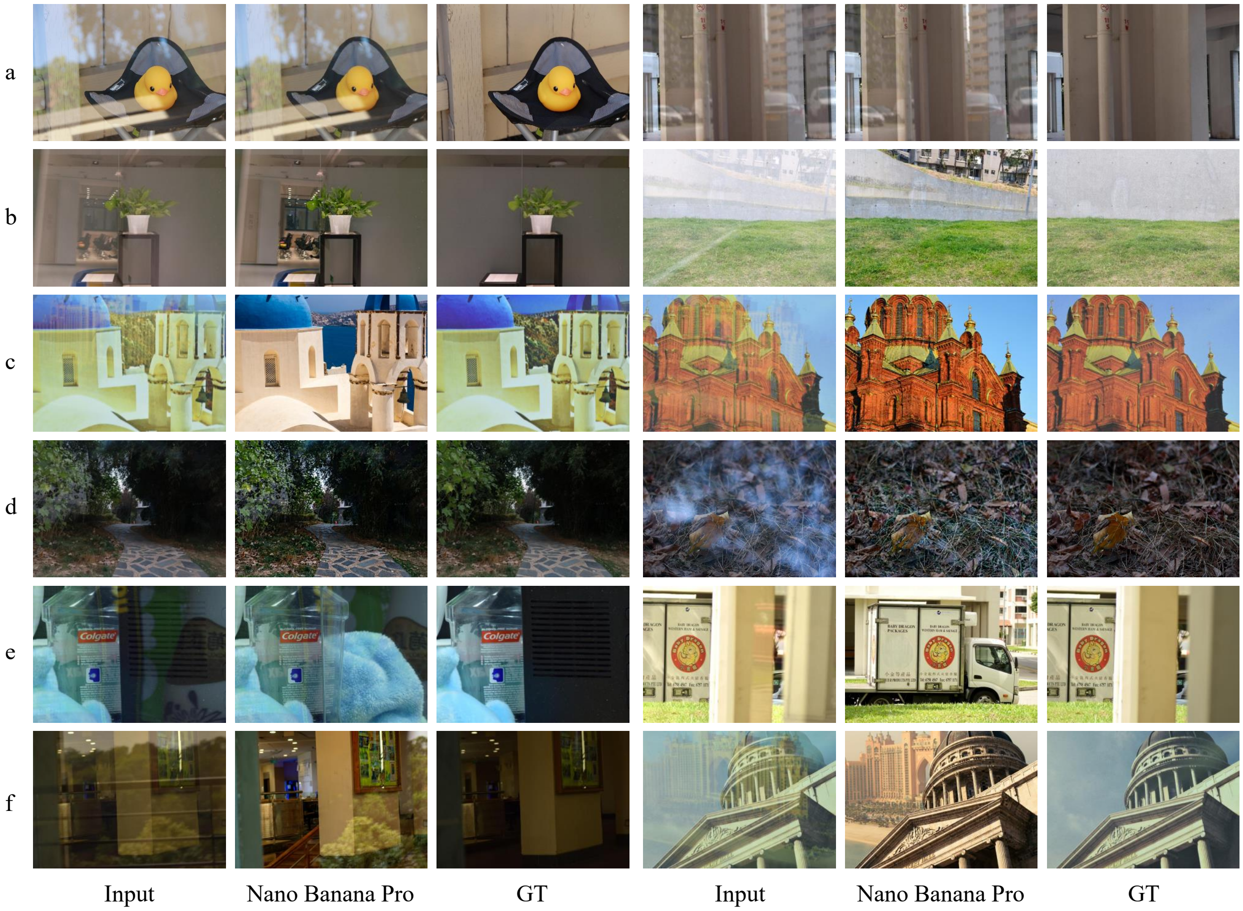


**Figure 22** Qualitative analysis of limitations and failure cases. We present typical examples where the proposed method yields suboptimal results, categorized by specific degradation types: (a) incomplete reflection removal due to strong intensity; (b) erroneous enhancement of reflection artifacts mistaken for background details; (c) unintended color deviation in the transmission layer; (d) significant texture distortion; (e) structural deformation compared to the ground truth; and (f) complex scenarios exhibiting compound artifacts involving multiple aforementioned issues.

**(a) Incomplete Reflection Removal.** In these instances, the model fails to effectively decouple the reflection layer from the transmission layer, resulting in significant residual artifacts. This behavior likely stems from a conservative inference strategy induced by prompts emphasizing background preservation. When the reflection intensity is high or statistically similar to the background, the model tends to classify the reflection as intrinsic scene content to avoid over-erasing potential background details.

**(b) Erroneous Enhancement due to Semantic Ambiguity.** Despite explicit instructions to suppress reflections, the model occasionally misinterprets reflection artifacts as valid background elements and erroneously enhances them. This phenomenon highlights a limitation in current generative priors: the model is driven by semantic plausibility rather than physical layer separation. When a reflection (e.g., a light source or architectural reflection) aligns semantically with the background scene, the model prioritizes generating a "coherent" image without contradictions, thereby integrating the artifact as a strengthened feature.

**(c) Unintended Chromatic and Domain Shift.** This error is predominantly observed in the *Postcard* dataset, which features images of paintings or prints (e.g., urban or humanist subjects). The model struggles with domain ambiguity, failing to distinguish between a "picture of a scene" and a "real-world scene." Consequently, it attempts to "restore" the printed content as a realistic photograph, aggressively altering saturation, removing characteristic grain, or modifying illumination. This results in severe color deviations and stylistic inconsistencies compared to the ground truth.

**(d) Texture Fidelity Loss.** In scenarios containing high-frequency details, such as natural foliage or fabric textures (e.g., towels), the generative process often fails to maintain the original texture distribution. The output tends to exhibit either unnatural over-smoothing (loss of fine grain) or artificial sharpening (introduction of high-frequency noise), indicating a lack of fine-grained control in the texture reconstruction module.

**(e) Structural Hallucination and Deformation.** In rare but severe cases, the model breaks structural consistency, generating outputs that deviate geometrically from the input. This includes the hallucination of non-existent background structures or the removal of actual objects mistaken for reflections. Such failures represent a collapse of the conditioning mechanism, where the strong generative prior overrides the spatial constraints provided by the input image.

**(f) Compound Degradation.** A significant portion of low-scoring results exhibits a hybrid of the aforementioned failure modes. For example, an image may suffer from incomplete reflection removal while simultaneously undergoing a global color shift, or experience structural deformation alongside texture smoothing. These complex scenarios represent the most challenging cases for the current architecture.

## 10 Flare Removal

### 10.1 Background

Lens flare constitutes a fundamental optical phenomenon wherein intense incident light undergoes scattering and reflection within a camera's lens system, resulting in parasitic artifacts that degrade image quality. These artifacts manifest predominantly as two distinct categories: scattering flares and reflective flares. Beyond aesthetic degradation, these artifacts critically impair downstream computer vision applications, including stereo matching misestimation, optical flow corruption, and semantic segmentation misclassification, thereby posing substantial risks to safety-critical systems such as autonomous driving and aerial object tracking [64, 88, 164].

Contemporary flare removal methodologies have evolved from traditional detection-based approaches to sophisticated deep learning frameworks enabled by large-scale datasets. The Flare7K++ [25] dataset represents a pivotal advancement, providing 7,000 synthetic flares with 25 scattering and 10 reflective patterns, supplemented by 962 real-captured flare images (Flare-R) that capture complex degradation effects unattainable through simulation alone. Recent state-of-the-art approaches demonstrate remarkable progress: Deflare-Mamba [62] introduces the first State Space Model (SSM)-based architecture, employing a hierarchical U-shaped framework with local-enhanced selective scan mechanisms to maintain contextual consistency across global flare patterns and local scene details. Meanwhile, the MiAlgo AI team achieved top performance in the MIPI 2024 Nighttime Flare Removal Challenge through a Progressive Perception Diffusion Network (PPDN) [26], combining an IR-SDE diffusion module for comprehensive flare elimination with an AOT Block enhancement stage for detail recovery, employing a two-stage progressive strategy to improve visual quality.

Performance assessment is conducted on two complementary benchmark suites. The Flare7K++ [25] test set comprises 100 meticulously aligned 512×512-resolution real-world flare-corrupted/flare-free image pairs, with manual annotations delineating glare, streak, and light source regions to enable component-specific evaluation via G-PSNR and S-PSNR metrics. Additionally, the MIPI 2024 Challenge introduces FlareReal600 [26], a high-resolution dataset featuring 600 aligned training images, with validation and test sets each containing 50 pairs available in both 2K (1440×1920) and 4K (1774×3840) resolutions to facilitate comprehensive evaluation across different spatial scales. In this subsection, we will systematically evaluate the flare removal capability of the Nano Banana Pro model on these benchmarks. We will examine its effectiveness in eliminating various nighttime flare artifacts while maintaining the scene's semantic and photometric integrity across different image resolutions, thereby providing a reference for the community.

**Table 10** Quantitative comparisons of Nano Banana Pro and representative specialists on the Flare7K++ dataset.

| **Metric** | Input | Restormer [184] | Uformer [157] | Flare-level [27] | DeflareMamba [62] | NB Pro |
|---|---|---|---|---|---|---|
| PSNR↑ | 22.56 | 27.60 | 27.63 | 27.05 | 26.06 | 24.92 |
| SSIM↑ | 0.857 | 0.897 | 0.894 | 0.901 | 0.898 | 0.844 |

**Table 11** Quantitative comparisons of Nano Banana Pro and representative specialists on the FlareReal600 dataset.

| **Metric** | PPDN [26] | NB Pro(2K) | NB Pro(4K) |
|---|---|---|---|
| LPIPS [196]↓ | 0.143 | 0.287 | 0.361 |
| PSNR↑ | 22.15 | 19.07 | 18.32 |
| SSIM [159]↑ | 0.708 | 0.496 | 0.424 |

## 10.2 Qualitative and Quantitative Results

To comprehensively assess the capabilities of Nano Banana Pro, we organized our experiments into quantitative evaluation and qualitative analysis.

**Quantitative Evaluation.** We first evaluated the model on the Flare7K++ dataset configured to an output resolution of 1K. Performance was measured using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [159]. As shown in Tab. 10, we compared Nano Banana Pro against state-of-the-art methods trained on the same dataset, including Restormer [184], Uformer [157], and DeflareMamba [62]. Subsequently, we extended our evaluation to the FlareReal600 dataset, processing images at their native resolutions to output 2K and 4K results. In addition to PSNR and SSIM, Learned Perceptual Image Patch Similarity (LPIPS) [196] was included to assess perceptual quality. Tab. 11 benchmarks our results against the MIPI 2024 Challenge champion, MiAlgo AI. Note that while the challenge metrics were derived from an unpublished test set, our evaluation utilized the publicly available validation set. We observed two notable quantitative trends: 1) Resolution Impact: Performance metrics generally decline as output resolution increases. 2) Brightness Sensitivity: On the high-resolution FlareReal600 dataset, higher image brightness results in degraded metrics. However, this trend is not evident in the lower-resolution Flare7K++ dataset.

**Qualitative Analysis.** Visual comparisons in Fig. 23 reveal a distinct dichotomy in the model's performance:

**1. Visual Superiority vs. Stochastic Instability**: On optimal inputs, Nano Banana Pro demonstrates exceptional deflaring capabilities, often surpassing SOTA methods in detail restoration. However, this advantage is compromised by the inherent stochasticity of diffusion models. The model exhibits significant variance and is prone to semantic hallucinations—such as generating unrelated content, suppressing valid light sources, or erroneously illuminating inactive bulbs. While prompt engineering offers partial mitigation, it fails to guarantee the deterministic reliability required for industrial deployment.

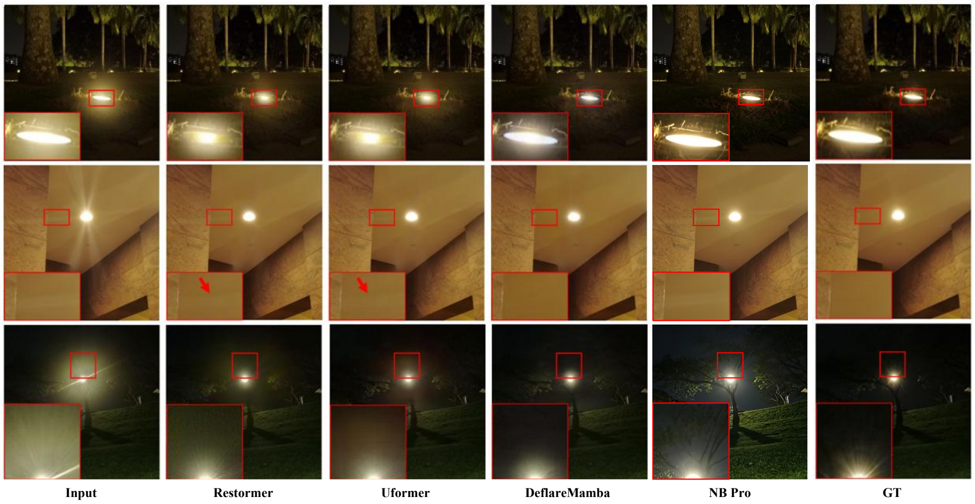


**Figure 23** Qualitative comparison of flare removal results on the Flare7K++ dataset. Nana Banana Pro can preserve image details near light sources and achieves clean removal of streak artifacts. However, it may introduce some brightness changes, as shown in the third row.

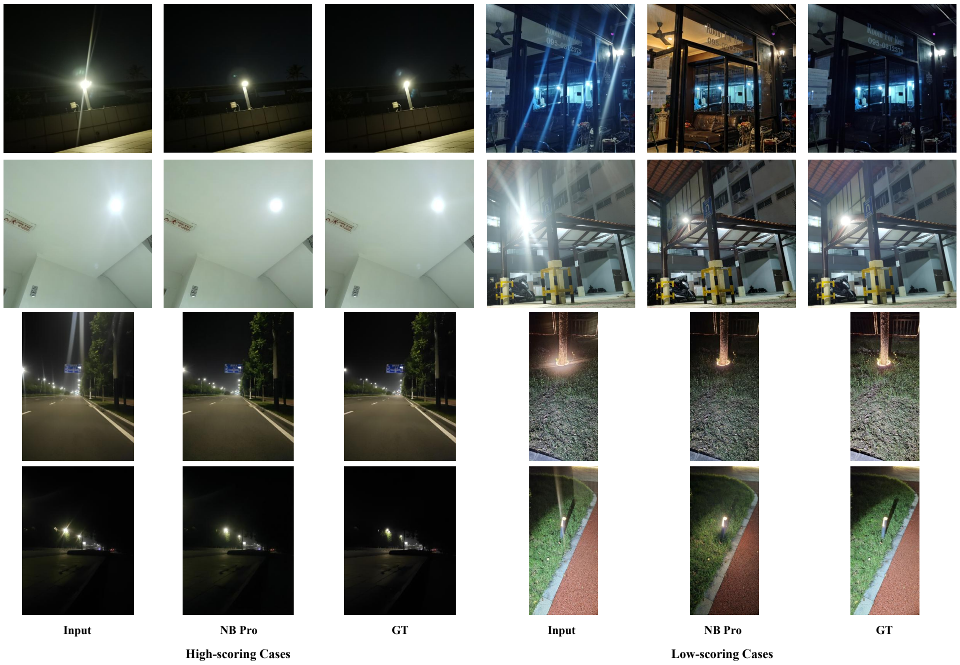


**Figure 24** Examples of some high-scoring and low-scoring samples from Nano Banana Pro on the Flare7K++ dataset. It can be observed that low-scoring samples sometimes appear visually satisfactory, yet their quantitative metric scores may be lower due to certain discrepancies in brightness or color compared to the ground truth.

**2. Perceptual-Metric Misalignment**: As illustrated in Fig. 24, we observe a notable divergence between quantitative metrics and perceptual quality. Instances with low scores sometimes retain high visual fidelity, suggesting that pixel-level metrics may not fully capture the perceptual advantages of generative reconstruction.

In conclusion, Nano Banana Pro exhibits a “high ceiling, low floor” characteristic. While it possesses the generative potential to outperform traditional regression-based methods in perceptual quality, it currently sacrifices the stability and consistency essential for robust image restoration.

# Image Enhancement

## 11 Low Light Image Enhancement

### 11.1 Background

Low-light image enhancement [11, 160, 179] aims to recover visually pleasing images from photographs captured under insufficient illumination. This task presents significant challenges, including brightness adjustment [14], noise suppression [54], and color restoration, all while preserving structural details and semantic content. Traditional approaches to this problem have relied on handcrafted priors or supervised deep learning methods trained explicitly on paired low-light and normal-light images.

Recent advances in unified multimodal models [100], which jointly handle image understanding and generation within a single framework, raise an intriguing question: can such models perform low-light enhancement in a zero-shot manner, leveraging their broad visual and semantic knowledge without task-specific training? In this chapter, we evaluate Nano Banana Pro, a unified generation and understanding multimodal model, on the low-light image enhancement task. Our evaluation spans three widely-used benchmarks, LOLv1 [160], LOLv2-real [179], and SICE [11], then we compare Nano Banana Pro's zero-shot performance against state-of-the-art supervised and unsupervised methods.

### 11.2 Experiment Setup

**Datasets.** We conduct experiments on three established low-light enhancement benchmarks. LOLv1 [160] contains 485 training pairs and 15 testing pairs of real-world low-light and normal-light images captured by adjusting camera exposure time and ISO. LOLv2-real [179] extends this with 689 training pairs and 100 testing pairs, featuring more diverse indoor and outdoor scenes. SICE [11] is a larger-scale dataset containing multi-exposure sequences, from which low-light and reference pairs are constructed; its test set comprises a more diverse range of scenes and illumination conditions.

**Evaluation Metrics.** we adopt two standard full-reference image quality metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Higher values indicate better reconstruction quality relative to the ground-truth normal-light images.

**Comparison Methods.** We compare Nano Banana Pro against several representative low-light enhancement methods spanning different paradigms: ZeroDCE [44] (zero-reference learning), RUAS [97] (architecture search-based), LLFlow [154](normalizing flow-based), LLFormer [149] (transformer-based), GSAD [54] (diffusion-based), and Quadprior [151] (diffusion prior-based). These methods represent the current state of the art in supervised and unsupervised low-light enhancement.

**Nano Banana Pro Configuration.** Nano Banana Pro is evaluated in a zero-shot setting without any fine-tuning on low-light enhancement data [11, 160, 179]. We provide the model with the following natural language instruction: "This is a low-light image, please turn this image into a normal image while keeping other elements unchanged." No additional inference-time configurations or post-processing steps are applied.

### 11.3 Qualitative and Quantitative Results

Tab. 12 presents the quantitative comparison across all three benchmarks. On LOLv1 [160] and LOLv2-real [179], Nano Banana Pro's zero-shot performance falls considerably short of the state-of-the-art supervised methods. The gap is particularly pronounced on LOLv2-real [179], where the PSNR of 15.661 dB and SSIM of 0.537 lag behind leading methods by a substantial margin. This suggests that without task-specific training, the model struggles to consistently produce enhancements that align with the ground-truth references in these benchmarks. Interestingly, on the SICE [11] dataset, Nano Banana Pro achieves slightly higher metrics than several comparison methods, demonstrating competitive zero-shot performance on this more challenging and diverse benchmark.

Fig. 25 presents representative visual comparisons across these three datasets [11, 160, 179]. Nano Banana Pro produces visually reasonable enhancements in many cases, successfully brightening dark regions and

**Table 12** Quantitative comparisons on the LOL and SICE datasets. The best results are highlighted by **black bold.**

| Methods | Color Model | LOLv1 [160] | | | LOLv2-Real [179] | | | SICE [11] | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| RetinexNet [160] | Retinex | 18.915 | 0.427 | 0.470 | 16.097 | 0.401 | 0.543 | 12.424 | 0.613 | - |
| KinD [199] | Retinex | 23.018 | 0.843 | 0.156 | 17.544 | 0.669 | 0.375 | - | - | - |
| ZeroDCE [44] | RGB | 21.880 | 0.640 | 0.335 | 16.059 | 0.580 | 0.313 | 12.452 | **0.639** | - |
| RUAS [97] | Retinex | 18.654 | 0.518 | 0.270 | 15.326 | 0.488 | 0.310 | 8.656 | 0.494 | - |
| LLFlow [154] | RGB | 24.998 | 0.871 | 0.117 | 17.433 | 0.831 | 0.176 | 12.737 | 0.617 | - |
| EnlightenGAN [68] | RGB | 20.003 | 0.691 | 0.317 | 18.230 | 0.617 | 0.309 | - | - | - |
| SNR-AW [173] | SNR+RGB | 26.716 | 0.851 | 0.152 | 21.480 | **0.849** | 0.163 | - | - | - |
| Bread [48] | YCbCr | 25.299 | 0.847 | 0.155 | 20.830 | 0.847 | 0.174 | - | - | - |
| PairLIE [37] | Retinex | 23.526 | 0.755 | 0.248 | 19.085 | 0.778 | 0.317 | - | - | - |
| LLFormer [149] | RGB | 25.278 | 0.823 | 0.167 | 20.856 | 0.792 | 0.211 | - | - | - |
| RetinexFormer [14] | Retinex | 27.140 | 0.850 | 0.129 | **22.794** | 0.840 | 0.171 | - | - | - |
| GSAD [54] | Retinex | **27.605** | **0.876** | **0.092** | 20.153 | 0.846 | **0.113** | - | - | - |
| QuadPrior [151] | Kubelka-Munk | 22.849 | 0.800 | 0.201 | 20.592 | 0.811 | 0.202 | - | - | - |
| **Nano Banana Pro** | - | **18.496** | **0.684** | **0.481** | **15.661** | **0.537** | **0.465** | **14.081** | **0.493** | - |

revealing scene content. However, the model exhibits inconsistent brightness control: in the first row, it tends to overexpose bright regions, while in others, it insufficiently enhances dark areas, leaving the output still underexposed. This inconsistency likely stems from the model's reliance on general visual priors rather than explicit illumination modeling. Notably, Nano Banana Pro does not introduce visible artifacts such as color distortion, halo effects, or structural corruption, which is a common failure mode of some enhancement methods. Texture preservation remains comparable to other approaches, with fine details in enhanced regions generally retained. The absence of artifacts suggests that the model's generative capabilities are well-regularized, even when applied to out-of-distribution tasks like low-light enhancement.

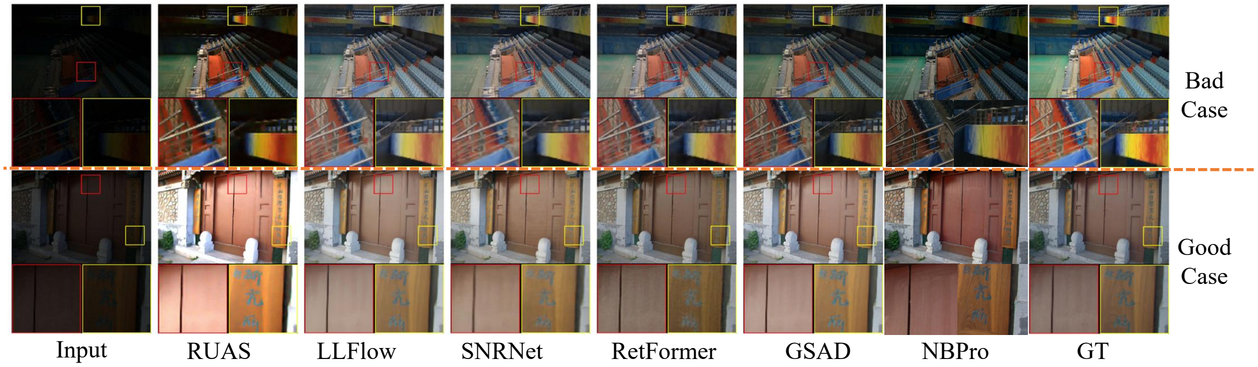


**Figure 25** Visual comparison examples of Nano Banana Pro and several representative specialists. The first row shows its shortcomings in brightness consistency, while the second row shows its superiority in detail preservation.

## 11.4 Analysis

The evaluation results reveal a nuanced picture of Nano Banana Pro's zero-shot capabilities for low-light image enhancement. In this section, we provide an in-depth analysis of the observed performance patterns, examine potential underlying causes, and discuss broader implications for applying unified multimodal models to image restoration tasks. A notable positive finding is that Nano Banana Pro does not introduce visible artifacts such as color distortion, halo effects around high-contrast edges, amplified noise, or structural corruption. This is significant because artifact introduction is a common failure mode of enhancement methods, particularly those based on generative models. The absence of artifacts may partially explain the moderate PSNR/SSIM scores. More aggressive enhancement methods might achieve higher metrics on average by pushing brightness and contrast more strongly, but at the cost of occasional artifacts. Nano Banana Pro's conservative approach avoids such failures but may sacrifice peak performance. For practical applications where artifact-free outputs are critical, this trade-off may be acceptable. However, the performance gap on standard benchmarks indicates that zero-shot application is not yet competitive with task-specific methods. The lack of explicit illumination modeling, sensitivity to prompt formulation, and inability to match benchmark-specific ground-truth definitions all limit current performance. Several avenues could potentially improve performance: 1) prompt engineering to provide more specific enhancement guidance; 2) few-shot learning with example image pairs to calibrate the model's enhancement behavior; 3) lightweight fine-tuning or adapter-based adaptation to inject task-specific knowledge while preserving general capabilities; and 4) hybrid approaches that combine unified models' semantic understanding with task-specific enhancement modules.

# 12 Underwater Image Enhancement

## 12.1 Background

Underwater image degradation is an inherent visual quality loss issue in marine environments. As light propagates through water, it undergoes selective absorption by water molecules, multiple scattering by suspended particles such as plankton and sediments, and complex illumination fluctuations, resulting in a series of characteristic image defects. These defects manifest primarily in three core types: color distortion, where red light rapidly attenuates in shallow waters, resulting in a blue-green color bias; contrast reduction stemming from haze-like obscuration caused by scattered light, significantly lowering the image signal-to-noise ratio; and texture blurring, a direct consequence of high-frequency edge information being obscured by scattered particles. Beyond compromising the visual presentation of underwater scenes, these degradations severely undermine the reliability of downstream computer vision tasks. This includes missed detections in underwater object recognition, biased seabed topography reconstruction, and misinterpreted biological behavior analysis. Consequently, it poses significant risks to safety-critical applications such as marine resource exploration, underwater cultural heritage archaeology, unmanned submersible navigation, and seabed infrastructure maintenance.

Traditional underwater image enhancement methods primarily rely on passive restoration, typically employing pixel-domain processing [5, 6, 39, 40, 65] or physical modeling [31, 38, 50, 75, 76, 197] to optimize images. However, these methods rely on manually designed rules and exhibit weak generalization capabilities in complex underwater scenes. Deep learning approaches for underwater image enhancement adopt a data-driven strategy, leveraging neural network architectures to learn end-to-end mapping relationships between degraded and clear images. They significantly outperform traditional methods in color correction and detail restoration. Diffusion model approaches for underwater image enhancement [137, 202] leverage progressive noise injection and reverse denoising mechanisms to effectively balance global tonal restoration with local texture preservation, emerging as one of the leading techniques in recent years. The UIEB [78] dataset serves as the benchmark for underwater image enhancement. This section utilizes its released 89-image challenge test set, which contains degraded real images and corresponding high-quality reference images to evaluate models' adaptability to complex scenarios. The LSUI [118] dataset is a large-scale real underwater image dataset. This section adopts the test set partitioning from WF-Diff [202], utilizing 400 images, covering diverse water bodies and target categories, making it suitable for evaluating model generalization. The U45 [79] dataset contains 45 reference-free real underwater images categorized into green cast, blue cast, and haze degradation types, simulating practical application scenarios requiring evaluation without reference images.

Performance evaluation is based on the above three datasets. In this subsection, we systematically assess the underwater image enhancement capabilities of the Nano Banana Pro model, focusing on its effectiveness in eliminating blue/green color casts and restoring blurred textures while preserving scene semantic integrity—such as biological morphology and artifact structure—and maintaining luminance consistency, including natural transitions between light and dark areas.

## 12.2 Quantitative Results

We conducted experiments using the Nano Banana Pro model configured to output 1K resolution across three datasets, quantitatively evaluating underwater image enhancement performance through reference-free metrics. The Underwater Image Quality Measure (UIQM) [117] comprehensively quantifies underwater image quality by integrating color richness, sharpness, and contrast. Underwater Color Image Quality Evaluation (UCIQE) [177]

**Table 13** Quantitative comparisons on UIEB, LUSI and U45 datasets. The best results are highlighted by **black bold**.

| Method | UIEB | | LUSI | | U45 | |
|---|---|---|---|---|---|---|
| | *UIQM*↑ | *UCIQE*↑ | *UIQM*↑ | *UCIQE*↑ | *UIQM*↑ | *UCIQE*↑ |
| UWCNN [78] | 3.8325 | 0.5552 | 4.1699 | 0.5453 | 4.387 | 0.5622 |
| UIEC$^2$-Net [153] | 3.327 | 0.609 | 3.9833 | 0.5888 | **4.4293** | 0.6104 |
| U-Shape [118] | 3.332 | 0.5751 | 4.0334 | 0.574 | 4.3524 | 0.5856 |
| PUGAN [22] | 3.2163 | **0.6176** | 4.0417 | 0.5834 | 4.3377 | **0.6117** |
| DM-Water [137] | **3.8925** | 0.5994 | 4.0595 | 0.5883 | 4.1986 | 0.586 |
| WF-Diff [202] | 3.7388 | 0.5867 | 4.0308 | 0.5688 | 4.2193 | 0.5813 |
| NB Pro | 3.634 | 0.5899 | **4.2993** | **0.5961** | 4.3907 | 0.5899 |

addresses non-uniform color shifts and low contrast in underwater scenes by evaluating quality across standard deviation, luminance contrast, and saturation mean dimensions. Both metrics adapt to real-world unreferenced scenarios without requiring reference images. Results were compared against UWCNN [78], UIEC$^2$-Net [153], U-Shape [118], PUGAN [22], DM-water [137], and WF-Diff [202]. Relevant details are summarized in Tab. 13.

NB Pro demonstrates competitiveness on underwater image reference-free evaluation metrics UIQM and UCIQE that rivals existing mainstream UIE methods. On the UIEB dataset, NB Pro's gap with optimal results in reference-free metrics is relatively small. On the larger-scale LSUI dataset with more complex degradation scenarios, NB Pro achieves top performance in both UIQM and UCIQE metrics, fully validating its robustness in challenging environments. On the reference-free dataset U45, NB Pro achieves the best UIQM score and ranks third in UCIQE, demonstrating its practical value in real-world reference-free scenarios.

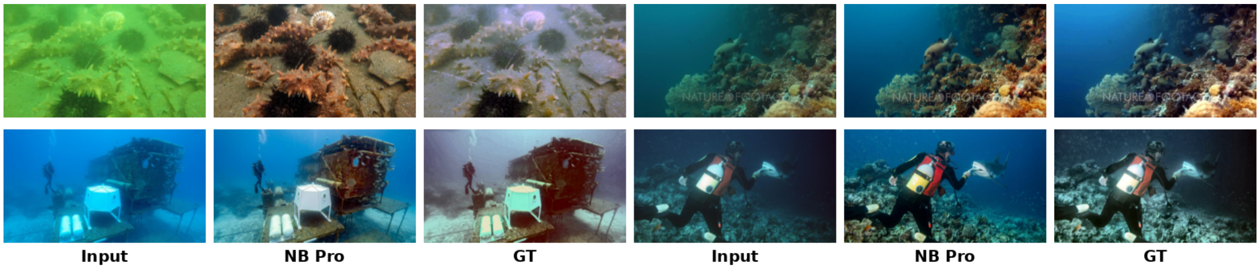


**Figure 26** Visualization examples for underwater image enhancement using NB Pro on the UIEB and LSUI datasets.

## 12.3 Qualitative Results

To intuitively present the underwater image enhancement outcomes of the Nano Banana (NB) Pro model, we provide visualizations of its processing results across the UIEB, LSUI, and U45 datasets, with comparisons to state-of-the-art baseline methods. Fig. 26 displays the visualization of exemplary cases for NB Pro in underwater image enhancement tasks on the UIEB and LSUI datasets. Fig. 27 presents the visualization of failed cases for NB Pro in underwater image enhancement tasks on the UIEB and LSUI datasets. Fig. 28 shows the visual comparison of processing results between NB Pro and other mainstream underwater image enhancement methods on the reference-free U45 dataset.

Qualitative experimental results demonstrate that for extreme degradation scenarios such as severe green color bias, severe blue color bias, insufficient illumination, and high turbidity, NB Pro can generate high-quality enhanced results without relying on GT images from paired training data. Even in scenarios with multiple severe degradations overlapping, its output images exhibit superior visual quality compared to GT images (as shown in Fig. 26). However, in mildly degraded scenarios with weaker color shifts, NB Pro struggles to effectively identify degradation features. The generated enhanced images exhibit minimal differences from the input images, resulting in visual quality inferior to GT images and failing to fully meet the core objective of underwater image enhancement (as shown in Fig. 27). This conclusion is further validated by visualization results from the reference-free dataset U45 (Fig. 28): In the first row depicting a diver scene with high turbidity and severe green color cast degradation, NB Pro's enhancement significantly outperforms other comparison methods, completely eliminating the color cast while removing foggy blur; In the second row of underwater scenes with blue color cast, NB Pro also performed comparably to existing mainstream methods; however, in the third row of slightly degraded underwater plant scenes, NB Pro's restoration results were relatively poor among all comparison methods, failing to effectively optimize image details and contrast.

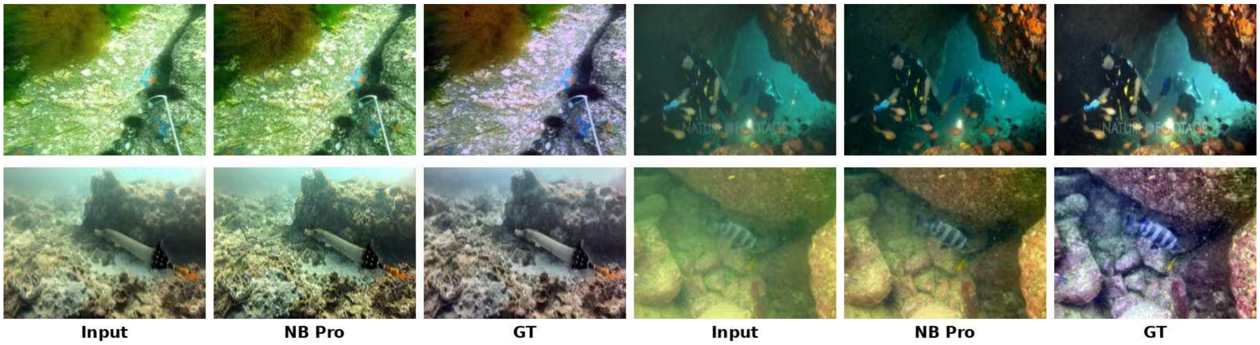


**Figure 27** Visualization of failed cases in underwater image enhancement for NB Pro on the UIEB and LSUI datasets.

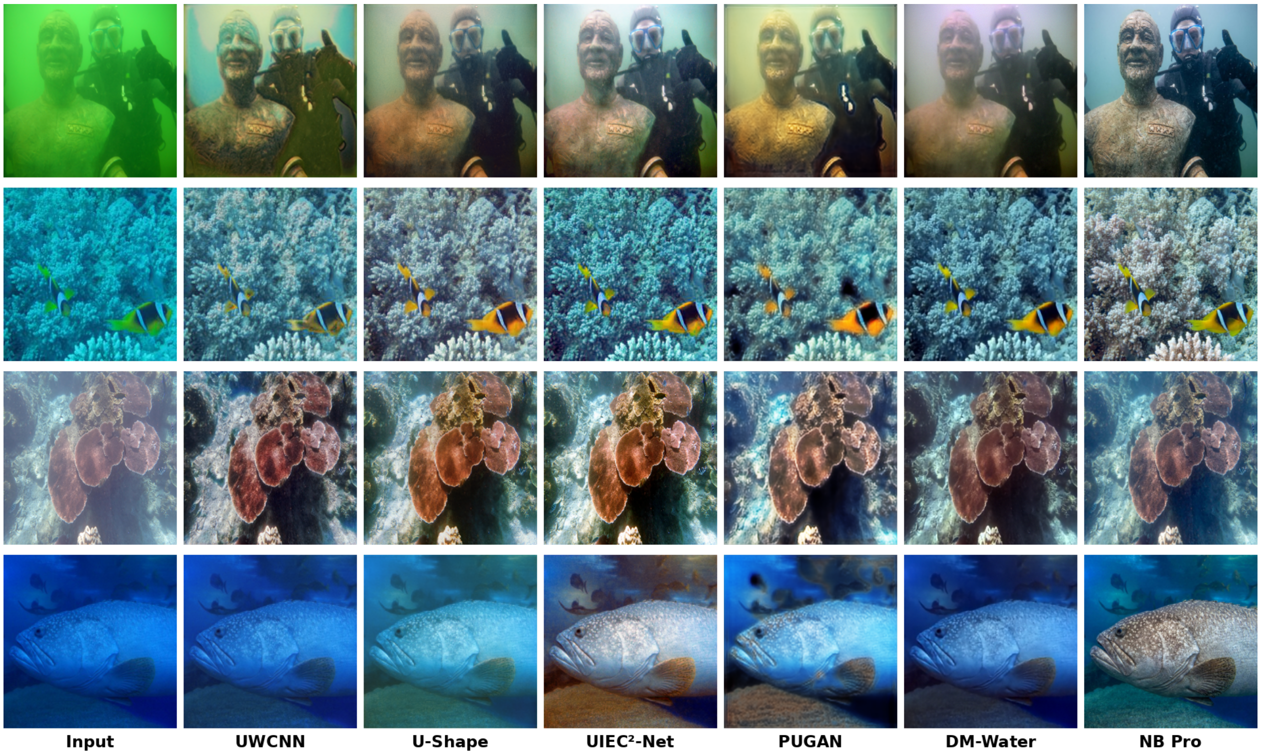


**Figure 28** Visual comparison of processing results between NB Pro and other underwater image enhancement methods on the U45 dataset without reference.

## 12.4 Analyses

This study conducted a comprehensive evaluation across multiple benchmark datasets representing both synthetic and real underwater environments. Results demonstrate that NB Pro offers a unique paradigm for underwater image enhancement. Its core characteristic is a distinct trade-off between robust perceptual recovery in complex environments and precise pixel-level fidelity in benign conditions.

Quantitative analysis demonstrates that NB Pro excels in the absence of reference images, achieving state-of-the-art results on large-scale complex datasets like LSUI and real-world datasets such as U45. This indicates that when confronted with severely degraded images, the model efficiently generates images with perceptual clarity, rich color saturation, and optimal contrast. Existing underwater ground truth images often exhibit residual degradation or artifacts. NB Pro's generative freedom enables it to surpass the visual quality of these reference images, consistent with its performance in the no-reference evaluation.

From a qualitative perspective, NB Pro demonstrates unique strengths in handling extremely degraded scenarios. Visual examples prove that NB Pro can successfully reconstruct scenes severely affected by green/blue color casts, low light, and high turbidity. It can simultaneously synthesize clear details and correct severe color shifts, sometimes producing results that visually outperform the ground truth images. This indicates NB Pro possesses a strong understanding of underwater degradation and the potential to reverse such degradation.

Despite these strengths in high-intensity restoration tasks, the model exhibits instability in mildly degraded scenarios. For scenes with only slight color shifts or minor turbidity, its enhancement effects are suboptimal, reflecting NB Pro's sensitivity limitations. When degradation signals are weak, the model struggles to identify or localize degraded features, failing to trigger necessary optimizations for fine details and contrast. Future research should focus on enhancing NB Pro's adaptability across the full spectrum of degradation levels. This will ensure it achieves both refined enhancement in mildly blurred environments and thorough reconstruction in severely degraded scenarios, delivering outstanding results in both cases.

# 13 HDR Imaging

## 13.1 Background

High Dynamic Range reconstruction technology aims to address the core limitations of traditional imaging systems. Conventional imaging systems fail to capture the complete spectrum of light intensity information in real-world scenes leading to irreversible loss of highlight details namely overexposed regions and shadow information namely underexposed regions in low dynamic range images. This inherent limitation not only degrades visual quality but also severely impairs the performance of downstream computer vision tasks such as misaligned object detection blurred scene parsing and inaccurate depth estimation posing substantial risks to safety-critical applications including intelligent surveillance environmental perception for autonomous driving and aerial scene analysis.

HDR reconstruction methods have evolved steadily forming a sophisticated technical framework anchored in large-scale datasets and driven by deep learning. Notably the MIT FiveK[10] dataset serves as a classic benchmark for tone mapping and image enhancement offering large-scale professional-grade HDR-LDR image pairs. It includes 5,000 image pairs 4,500 for model training augmented via random cropping flipping rotation and other techniques and 500 for validation and testing. The dataset covers diverse real-world scenarios from indoor and outdoor environments to low-light and high-contrast scenes and supplies professional-grade HDR reference images with exposure parameters and semantic annotations. This enables the reproduction of real-world dynamic range variations and meets evaluation requirements for algorithms focusing on global tone adjustment and local detail preservation. The HDR+[141] dataset is a large-scale benchmark for real-world HDR imaging focusing on the actual shooting conditions of consumer cameras. It contains over 100,000 RAW image sequences captured by consumer cameras each paired with aligned HDR ground truth images. Its advantages include capturing natural noise patterns sensor characteristics and complex lighting variations such as backlighting and high contrast simulating real-world tone mapping demands while retaining resolution suitable for practical use.

Performance evaluation leverages two benchmark datasets. The FiveK dataset includes 500 aligned LDR and HDR image pairs and the HDR+ benchmark emphasizing real-world applicability includes 250 test image pairs. Evaluations use quantitative metrics including PSNR SSIM and LPIPS complemented by subjective assessments of naturalness and detail fidelity to comprehensively measure model performance. In this section we systematically evaluate the Nano Banana Pro model's HDR reconstruction performance on 480p images using the aforementioned benchmarks. The assessment prioritizes verifying the model's capacity to recover highlight and shadow details across diverse scenes and spatial scales while ensuring semantic consistency and photometric integrity. These results offer valuable reference data for the research community.

## 13.2 Quantitative Results

**Table 14** Quantitative comparison on HDR+ and MIT-FiveK datasets. The best results are highlighted by **black bold**.

| **Method** | **HDR+ (480p)** | | | | **MIT-FiveK (480p)** | | | |
|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | $\triangle E\downarrow$ | PSNR↑ | SSIM↑ | LPIPS↓ | $\triangle E\downarrow$ |
| UPE[147] | 23.33 | 0.852 | 0.150 | 7.68 | 21.82 | 0.839 | 0.136 | 9.16 |
| HDRNet[41] | 24.15 | 0.845 | 0.110 | 7.15 | 23.31 | 0.881 | 0.075 | 7.73 |
| CSRNet[49] | 23.72 | 0.862 | 0.104 | 6.67 | 25.31 | 0.909 | **0.052** | 6.17 |
| DeepLPF[113] | 25.73 | 0.904 | 0.073 | 6.05 | 24.97 | 0.897 | 0.061 | 6.22 |
| LUT[185] | 23.29 | 0.855 | 0.117 | 7.16 | 25.10 | 0.902 | 0.059 | 6.10 |
| sLUT[148] | 26.13 | 0.901 | 0.069 | 5.34 | 24.67 | 0.896 | 0.059 | 6.39 |
| CLUT[188] | 26.05 | 0.892 | 0.088 | 5.57 | 24.94 | 0.898 | 0.058 | 6.71 |
| LLF-LUT++[187] | **28.43** | **0.924** | **0.056** | **4.54** | **26.06** | **0.912** | 0.054 | **4.93** |
| **NB Pro** | **14.24** | **0.467** | **0.221** | **19.82** | **19.20** | **0.639** | **0.133** | **11.14** |

To comprehensively evaluate Nano Banana Pro's performance in HDR tasks, we quantitatively compared it against a range of advanced traditional and deep learning-based image enhancement methods. To ensure fair comparison, all images were downsampled to 480p resolution for evaluation. We employed four standard metrics: PSNR and SSIM to evaluate perceptual similarity, LPIPS to assess visual similarity, and $\triangle E$ to quantify color differences. Results are shown in Tab. 14. NB Pro significantly underperformed against the comparison methods. On the HDR+ dataset, NB Pro achieved lower PSNR and SSIM than the optimal method, while also exhibiting poorer LPIPS and $\triangle E$ values. On the MIT-FiveK dataset, although its PSNR and SSIM improved, they still lagged significantly behind the optimal method. This result clearly indicates that under the standard full-reference evaluation framework, which prioritizes pixel-level accurate reconstruction and color fidelity, NB Pro's generated results exhibit systematic deviations from professionally enhanced or color-graded reference images.

NB Pro fundamentally differs from traditional HDR/enhancement models optimized for specific imaging scenarios. The latter typically undergo end-to-end training directly on paired LDR-reference images, targeting minimization of pixel-level loss, thus inherently excelling in metrics like PSNR and SSIM. In contrast, NB Pro's generation process prioritizes semantic coherence and overall visual appeal. Its outputs can be viewed as reconstructions of the input image rather than strict pixel-to-pixel mappings. Consequently, generated images may exhibit deviations in luminance distribution, local contrast, and even color style compared to reference images, leading to comprehensive score reductions across full-reference metrics. Notably, on the LPIPS metric, NB Pro's performance on the MIT-FiveK dataset remains behind but shows a narrowed gap compared to pixel-level metrics. This suggests its outputs may retain some similarity to reference images at higher-level semantic features, while low-level pixel arrangements have been significantly altered.

## 13.3 Qualitative Results

To visually evaluate the performance and potential limitations of NB Pro in HDR imaging tasks, this study conducted qualitative visualization experiments using representative scenes from the HDR+ and MIT-FiveK datasets. The experiments cover three core scenarios: conventional lighting, low-light with minimal detail, and complex dense textures. The corresponding results are presented in Fig. 29, Fig. 30, and Fig. 31, respectively.

Fig. 29 presents an illustrative case of NB Pro in HDR imaging tasks. Comparing the low-dynamic-range input image, the true high-dynamic-range reference image, and the NB Pro generated result reveals that in conventionally lit scenes, such as moderately bright indoor settings or outdoor natural light environments, NB Pro effectively restores the scene's dynamic range and color gradation. The overall visual quality of the generated result approaches or even rivals the reference image, fully validating the model's effectiveness in fundamental HDR imaging tasks.

Fig. 30 further highlights texture anomalies in low-light, low-detail scenarios. When input low-dynamic-range images contain severe shadow areas with inherently weak texture or detail information, NB Pro tends to exhibit two typical defects. One type involves detail loss, such as in the third case where the model's output lacks the wall tile texture and sofa outline in the shadow regions of the input image. The second involves

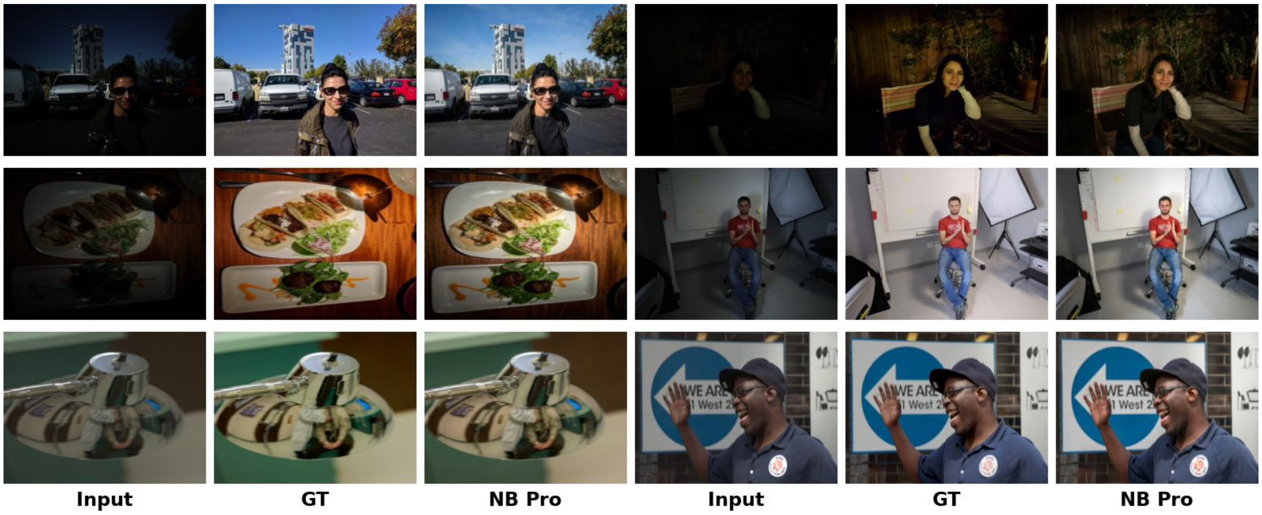


**Figure 29** Visualization of satisfactory exemplary cases for the HDR imaging task using NB Pro.

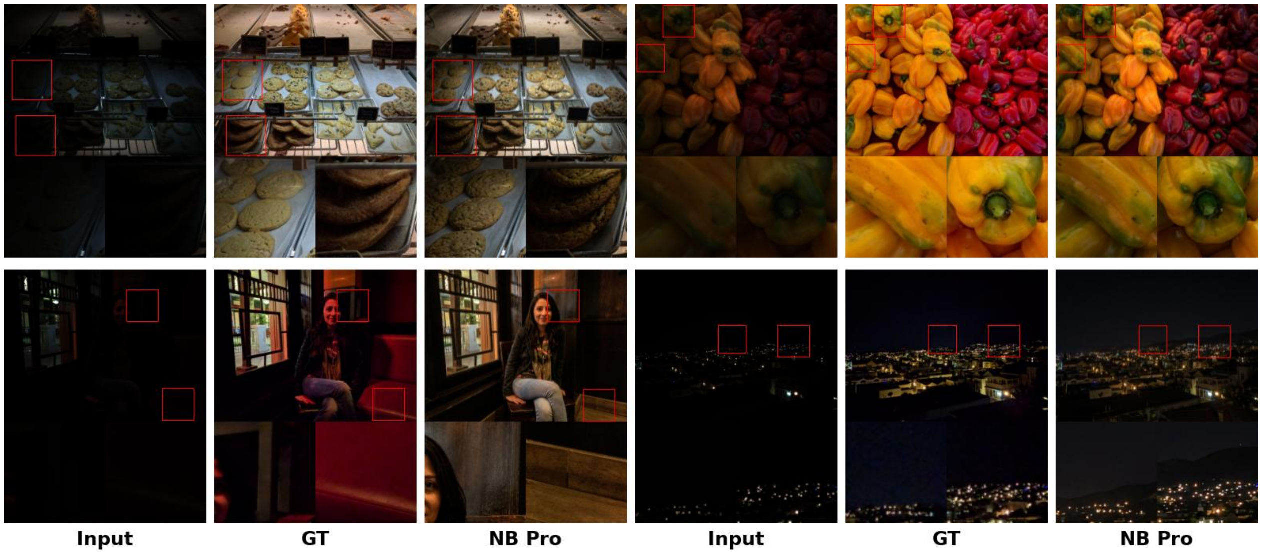


**Figure 30** Visualization of texture detail loss and artificial addition cases in HDR Imaging using NB Pro.

artificially redundant additions, such as the extra cookie surface texture generated in the first case that did not exist in the original scene. The second case failed to restore the vegetables' original colors and applied redundant diffuse rendering to the green spots. The fourth case generated mountain elements out of thin air in the house background. These issues stem from insufficient detail features in low-light scenes, causing the model's texture perception and generation logic to deviate, ultimately reducing detail restoration accuracy.

Fig. 31 shows evaluation results for complex, densely textured scenes. When input images contain fine, dense textures—such as fabric patterns, intricate vegetation textures, or architectural wall textures—NB Pro struggles to precisely balance texture preservation and enhancement scales. Generated results commonly exhibit visual artifacts from excessive sharpening, overly pronounced texture edges, localized artifacts, and even masking of the original texture's natural gradations. This phenomenon reflects the model's ongoing limitations in perceiving fine textures and controlling enhancement. In summary, qualitative experiments indicate that NB Pro delivers satisfactory visual effects in standard HDR imaging scenarios. However, in low-light, low-detail environments, the model tends to suffer from texture loss or redundant additions. In complex, densely textured scenes, the model exhibits a tendency toward excessive sharpening.

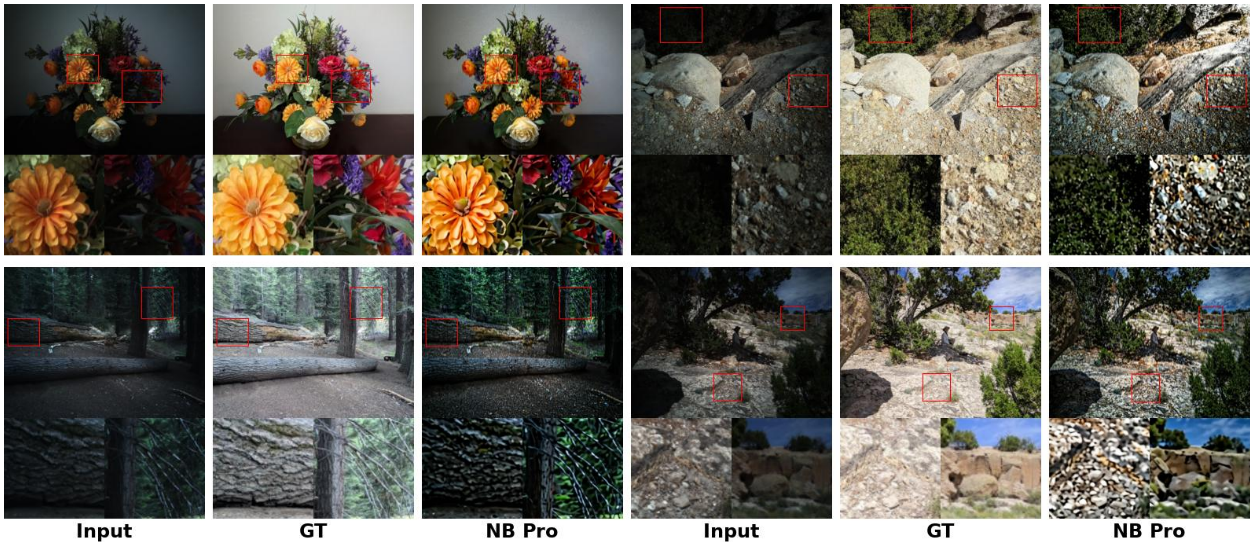


**Figure 31** Visualization of oversharpening cases in the HDR imaging task using NB Pro.

### 13.4 Analyses

The quantitative and qualitative evaluation results systematically reveal the comprehensive performance characteristics of NB Pro in HDR imaging tasks. Quantitatively, the model significantly lags behind mainstream HDR reconstruction methods in all reference evaluation metrics during 480p resolution tests on the HDR+ and MIT-FiveK datasets. However, the gap narrows for the LPIPS metric on the MIT-FiveK dataset, indicating that while its generated results exhibit significant pixel-level deviations, they maintain a degree of semantic consistency with reference images. Qualitatively, NB Pro achieves acceptable visual results in standard-illumination HDR scenes. However, it exhibits pronounced limitations in two typical scenarios: low-light environments with sparse details and complex, densely textured scenes. Specific issues include texture information loss, artificial redundant detail generation, color distortion, and over-sharpening artifacts.

The evaluation results clearly reveal three core deficiencies of NB Pro when applied to HDR imaging tasks: First, in low-light, low-detail scenes, the model exhibits weaknesses in perceiving and accurately restoring subtle texture features in shadow areas. The weak feature signals in these input shadows struggle to support precise reconstruction, leading to either the omission of shadow texture information or the generation of artificial redundant detail fill. Second, for fine-grained texture scenarios like fabric weaves and dense vegetation patterns, the model lacks adaptive control over texture preservation and enhancement scales. Over-sharpening not only disrupts natural texture gradation but also readily induces edge artifacts. Third, the model's color rendering mechanism lacks strong constraints on reference color distributions. In complex color scenes, it prioritizes visual harmony over precise target color reproduction, ultimately causing color distortion. Comprehensive experimental results indicate that NB Pro is only suitable for non-critical HDR imaging scenarios where pixel-level precision is less demanding and visual experience is paramount. It is unsuitable for safety-critical or professional-grade applications requiring stringent detail and color restoration accuracy. Issues such as texture loss, artificial redundant details, and color distortion may cause scene analysis bias or decision misjudgment, failing to meet the core requirements of such scenarios.

# Image Fusion

## 14 Multi-Focus Image Fusion

### 14.1 Introduction

In the computer vision and digital photography, acquiring fully clear images is crucial for subsequent analysis and processing. However, due to the limited depth of field (DoF) of camera lenses, a single shot cannot concurrently focus on all the objects at varying depths. Multi-focus image fusion (MFIF) provides an effective solution to this challenge, aiming to generate an AIF output from multiple images of the same scene with different focal points. This technique has found significant applications in various fields such as medical diagnosis and consumer electronics.

In recent years, deep learning has been widely applied to MFIF, typically falling into two categories: decision-based and reconstruction-based approaches. Decision-based methods learn a decision map by classifying pixels in the source images as either in-focus or out-of-focus, and then select the appropriate pixel to assemble the final composite [85, 99, 166]. Reconstruction-based methods employ end-to-end networks to directly extract features from source images and reconstruct the all-in-focus output [86, 110, 200]. However, the former methods normally struggle with focused-defocused boundaries whose focus attributes are ambiguous to distinguish, while the letter ones often suffer from detail loss in other regions away from the boundaries.

More recently, the rapid advancement of Generative Artificial Intelligence has opened new avenues for MFIF. Generative models learn the distribution from massive datasets, enabling them to understand and generate complex visual content and structures. Despite the potential for creative fusion, their application in pixel-precise tasks like MFIF remains under-explored. There is a critical need to verify whether a model designed for creativity can adhere to the strict fidelity requirements of fusion tasks without introducing hallucinations or losing spectral information. This report evaluates the performance of the newest generative model Nano Banana Pro in the task of MFIF, comparing its fusion quality with existing baseline methods using various metrics. The results aim to provide insights into its strengths and limitations in real-world applications.

### 14.2 Quantitative Results

**Table 15** Quantitative comparison on the Lytro and MFFW datasets. The best results are highlighted by **black bold**.

| Method | Lytro | | | | | | MFFW | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $NMI$ | $EN$ | $AG$ | $SF$ | $Q_Y$ | $Q_{CB}$ | $NMI$ | $EN$ | $AG$ | $SF$ | $Q_Y$ | $Q_{CB}$ |
| IFCNN [200] | 0.9388 | 7.5318 | 8.1473 | 19.3992 | 0.9518 | 0.7294 | 0.8206 | 7.1688 | 9.6849 | 22.0173 | 0.8715 | 0.6423 |
| SESF [105] | 1.6808 | 7.5322 | 8.1530 | 19.4251 | 0.9879 | 0.8064 | 1.0876 | 7.1850 | 9.8678 | 22.9291 | 0.9588 | 0.7418 |
| GACN [106] | 1.1723 | 7.5311 | 8.1245 | 19.3247 | 0.9878 | 0.8062 | 1.0820 | 7.1923 | 9.7328 | 22.2842 | 0.9273 | 0.7192 |
| GRFusion [81] | 1.1879 | 7.5329 | 8.1823 | 19.4539 | 0.9863 | 0.8071 | 1.1426 | 7.1711 | 9.8826 | 22.6520 | 0.9334 | 0.7223 |
| ZMFF[60] | 0.8795 | 7.5223 | 7.7771 | 18.8184 | 0.9321 | 0.6731 | 0.7711 | 7.1546 | 9.2144 | 21.3405 | 0.8474 | 0.6722 |
| MUFusion [21] | 0.8088 | **7.6093** | 8.1578 | 18.9240 | 0.8997 | 0.6819 | 0.7551 | 7.2026 | 8.9247 | 19.9974 | 0.8167 | 0.6205 |
| DB-MFIF [191] | 1.0573 | 7.5386 | 8.2415 | 19.5290 | 0.9637 | 0.7770 | 0.8699 | 7.1935 | 10.1346 | 23.0305 | 0.8663 | 0.6647 |
| MFFT [186] | 1.1533 | 7.5620 | **8.6089** | **21.3759** | 0.9523 | 0.7511 | 1.1310 | **7.2281** | 10.1971 | **24.5709** | 0.9336 | 0.6865 |
| DMANet [125] | 1.1897 | 7.5319 | 8.2059 | 19.5129 | 0.9853 | 0.8054 | 1.1513 | 7.1846 | 10.0948 | 23.2370 | 0.9506 | 0.7276 |
| MCCSR [209] | **1.1920** | 7.5329 | 8.1757 | 19.4392 | **0.9890** | **0.8084** | **1.1800** | 7.1688 | 9.7407 | 22.2989 | **0.9813** | **0.7517** |
| **NB Pro** | **0.7476** | **7.5267** | **8.2977** | **20.1044** | **0.7755** | **0.6603** | **0.6319** | **7.1823** | **10.2879** | **23.0223** | **0.5537** | **0.5638** |

We evaluate the performance of MFIF on four benchmark: Lytro [116], MFFW [172], MFI-WHU [190] and SIMIF [139]. The Lytro dataset contains 20 pairs of multi-focus images captured by a light field camera. The MFFW dataset includes 13 real image pairs with strong Defocus Spread Effect (DSE). The MFI-WHU dataset is constructed using Gaussian blur and decision maps, consists of a larger scale with 120 pairs. The SIMIF dataset is composed of 12 pairs of high-resolution images. Six popular objective metrics are employed for evaluation, including non-reference metrics, $EN$ [66], $AG$ [23], $SF$ [210], and Source-reference metrics $NMI$ [53], $Q_Y$ [175], $Q_{CB}$ [18]. These datasets and metrics provide a comprehensive assessment from multiple perspectives.

**Table 16** Quantitative comparison on the MFI-WHU and SIMIF datasets. The best results are in **black bold**.

| Method | MFI-WHU | | | | | | SIMIF | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $NMI$ | $EN$ | $AG$ | $SF$ | $Q_Y$ | $Q_{CB}$ | $NMI$ | $EN$ | $AG$ | $SF$ | $Q_Y$ | $Q_{CB}$ |
| IFCNN [200] | 0.8993 | 7.3285 | 11.3564 | 26.1798 | 0.9404 | 0.7367 | 1.0505 | 7.3897 | 6.9289 | 19.1344 | 0.9211 | 0.7364 |
| SESF [105] | 1.1878 | 7.3183 | 11.5399 | 26.5894 | 0.9855 | 0.8166 | 1.2995 | 7.3868 | 6.9931 | 19.4672 | 0.9807 | 0.8306 |
| GACN [106] | 1.2084 | 7.3127 | 11.4633 | 26.5089 | 0.9889 | **0.8241** | 1.3068 | 7.3802 | 6.9750 | 19.4241 | **0.9845** | **0.8328** |
| ZMFF[60] | 0.7028 | 7.2763 | 10.6727 | 24.7632 | 0.8387 | 0.6512 | 0.8302 | 7.3942 | 7.0007 | 19.2939 | 0.8325 | 0.6259 |
| GRFusion [81] | 1.2134 | 7.3216 | 11.6609 | 26.8362 | 0.9848 | 0.8212 | 1.2898 | 7.3917 | 7.0215 | 19.4954 | 0.9771 | 0.8267 |
| MUFusion [21] | 0.7449 | **7.3744** | 9.6015 | 21.4447 | 0.8608 | 0.6480 | 0.8876 | 7.3566 | 6.0633 | 16.3885 | 0.8387 | 0.6321 |
| DB-MFIF [191] | 1.0693 | 7.3250 | 11.6956 | 26.8336 | 0.9466 | 0.7818 | 1.1510 | 7.4073 | 7.1849 | 19.7543 | 0.9157 | 0.7612 |
| MFFT [186] | 1.1974 | 7.3358 | **11.7636** | **27.5736** | 0.9614 | 0.7648 | 1.2751 | 7.3866 | 7.1754 | **20.5918** | 0.9492 | 0.7724 |
| DMANet [125] | 1.2220 | 7.3158 | 11.6003 | 26.8034 | 0.9878 | 0.8222 | 1.3130 | 7.3834 | 7.0634 | 19.5735 | 0.9780 | 0.8270 |
| MCCSR [209] | **1.2246** | 7.3120 | 11.4162 | 26.4025 | **0.9891** | 0.8227 | **1.3151** | 7.3782 | 6.9240 | 19.3575 | 0.9835 | 0.8298 |
| **NB Pro** | **0.5923** | **7.2986** | **10.5142** | **24.1274** | **0.5610** | **0.5745** | **0.8042** | **7.4925** | **7.3561** | **19.5545** | **0.5112** | **0.5775** |

We compare the Nano Banana Pro (NB Pro) with 10 other state-of-the-art and representative MFIF methods, where ZMFF is a Zero-shot method, IFCNN and MUFusion are unsupervised methods, and the rest are supervised methods. According to the comparison results shown in the Tab. 32 and Tab. 33, NB Pro performs exceptionally well on non-reference metrics, achieving results that are close to or even surpassing the current state-of-the-art, indicating the high quality of the generated images themselves. Conversely, on source-reference metrics, NB Pro shows poorer performance, meeting the similar dilemma faced by previous Zero-shot and unsupervised methods. This indicates that during the fusion process, the model failed to adequately preserve consistency between the generated image and the source images in terms of aspects like gradients and structure; it exhibits excessive creativity at the expense of fidelity.

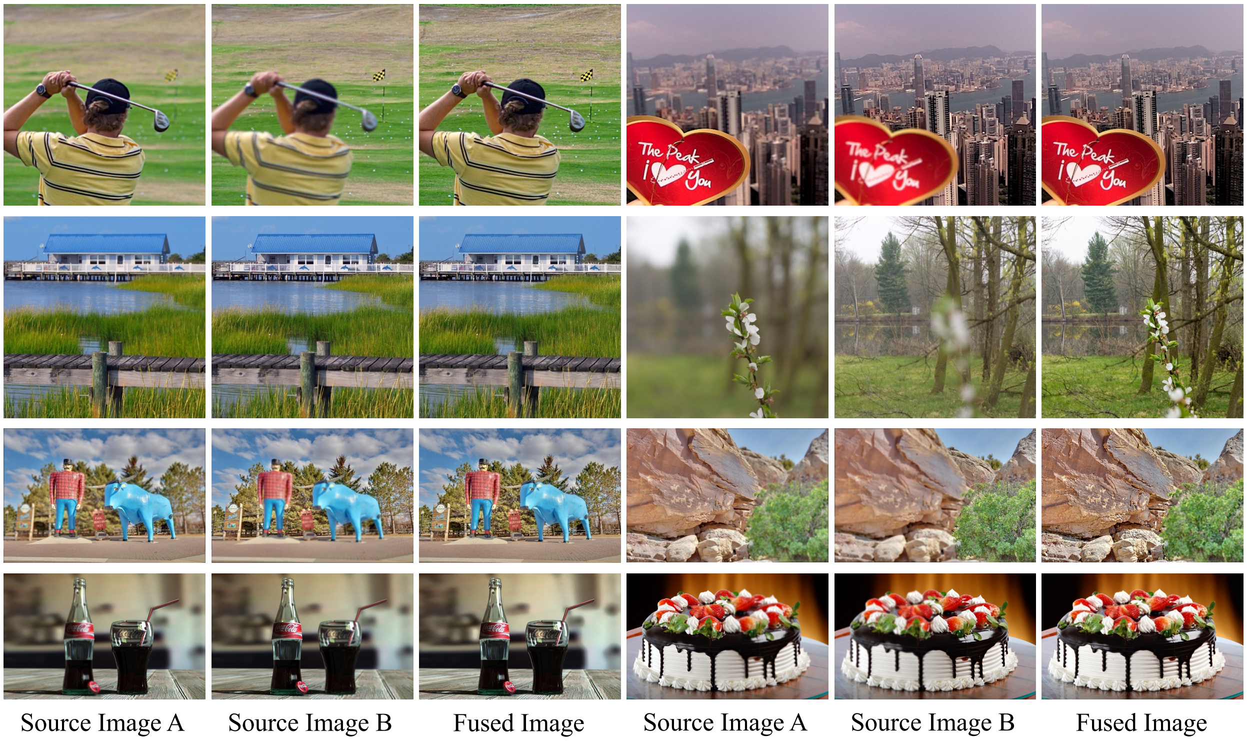


**Figure 32** Visualization of the fusion images generated by the Nano Banana Pro. The rows from top to bottom correspond to samples from the Lytro, MFFW, MFI-WHI and SIMIF datasets.

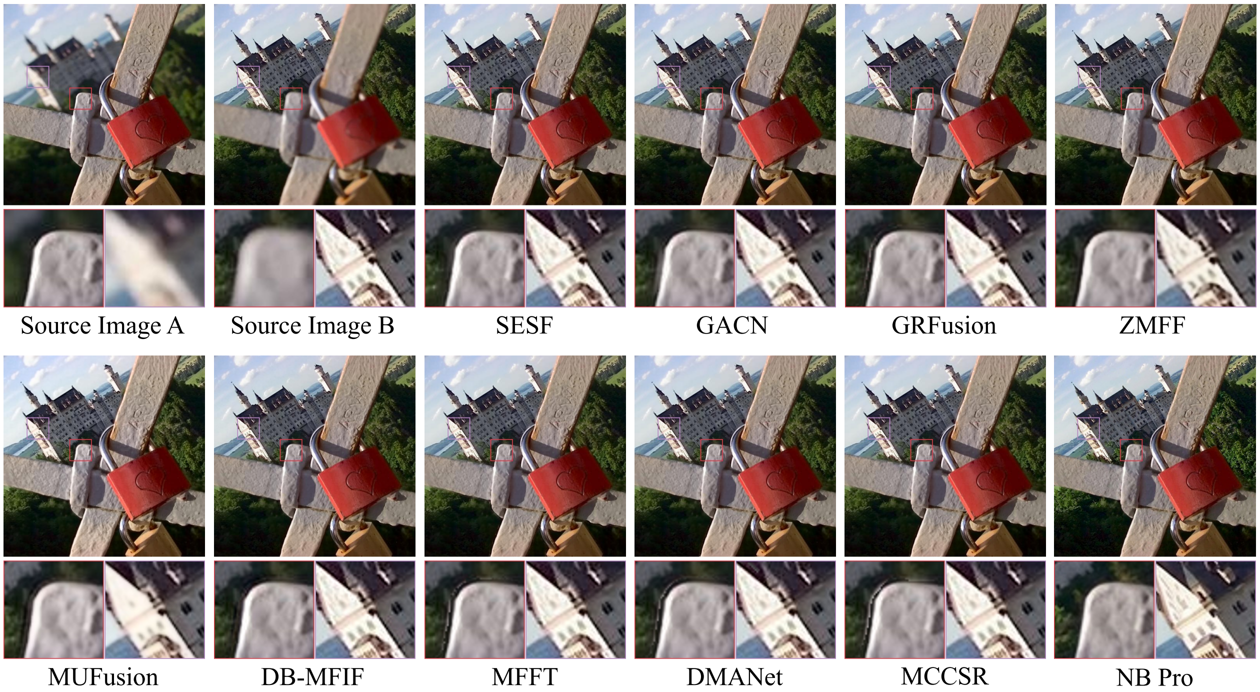


**Figure 33** Visualization of the fusion images of 'lock' sample from the Lytro dataset. Two enlarged views are shown to reveal critical details.

## 14.3 Qualitative Results

Fig. 32 illustrates a visualization of the fusion results generated by the Nano Banana Pro. The fusion performance varies across different samples, with some being successful and others subpar. For instance, the first example in the first row showcases a favorable fusion result, where the fused image not only preserves the focused foreground person and the background golf course from the source images but also achieves a seamless transition between different regions, effectively avoiding artifacts. Notably, it even recovers richer details in the lawn area—which originally had limited clarity—thereby enhancing the overall visual quality. Conversely, limitations are observed in some other cases. Specifically, in the second sample of the second row, the white petals at the base of the foreground plant remain blurred. Similarly, in the second sample of the third row, the grass in the bottom-right corner is still defocused. Given that sharp counterparts for these regions exist in the source images, this indicates that the model failed to correctly identify or localize the optimal in-focus regions during the fusion process.

To more intuitively verify the capability of NB Pro, Fig. 33 and Fig. 34 compares the fusion results against state-of-the-art MFIF approaches. In the 'lock' example, NB Pro achieves remarkably smooth transitions across regions without introducing artifacts near the lock head, while producing even sharper details on the house wall that was already in focus in the source images. In the 'coffee cup' example, due to the spread effect caused by defocus, some methods generate artifacts at the boundary between the two cup rims, and others produce dark ghosting along the cup wall. In contrast, NB Pro effectively overcomes these issues and delivers a visually pleasing fusion result.

## 14.4 Analyses

This comprehensive evaluation, conducted across four diverse benchmarks and benchmarked against ten state-of-the-art methods, establishes NB Pro as a paradigm shift in Multi-Focus Image Fusion. The results characterize a distinct trade-off between perceptual quality and signal fidelity.

Quantitative analysis reveals a significant performance divergence. NB Pro excels in non-reference metrics, generating images with exceptional clarity, textural richness, and visual appeal. Conversely, its performance

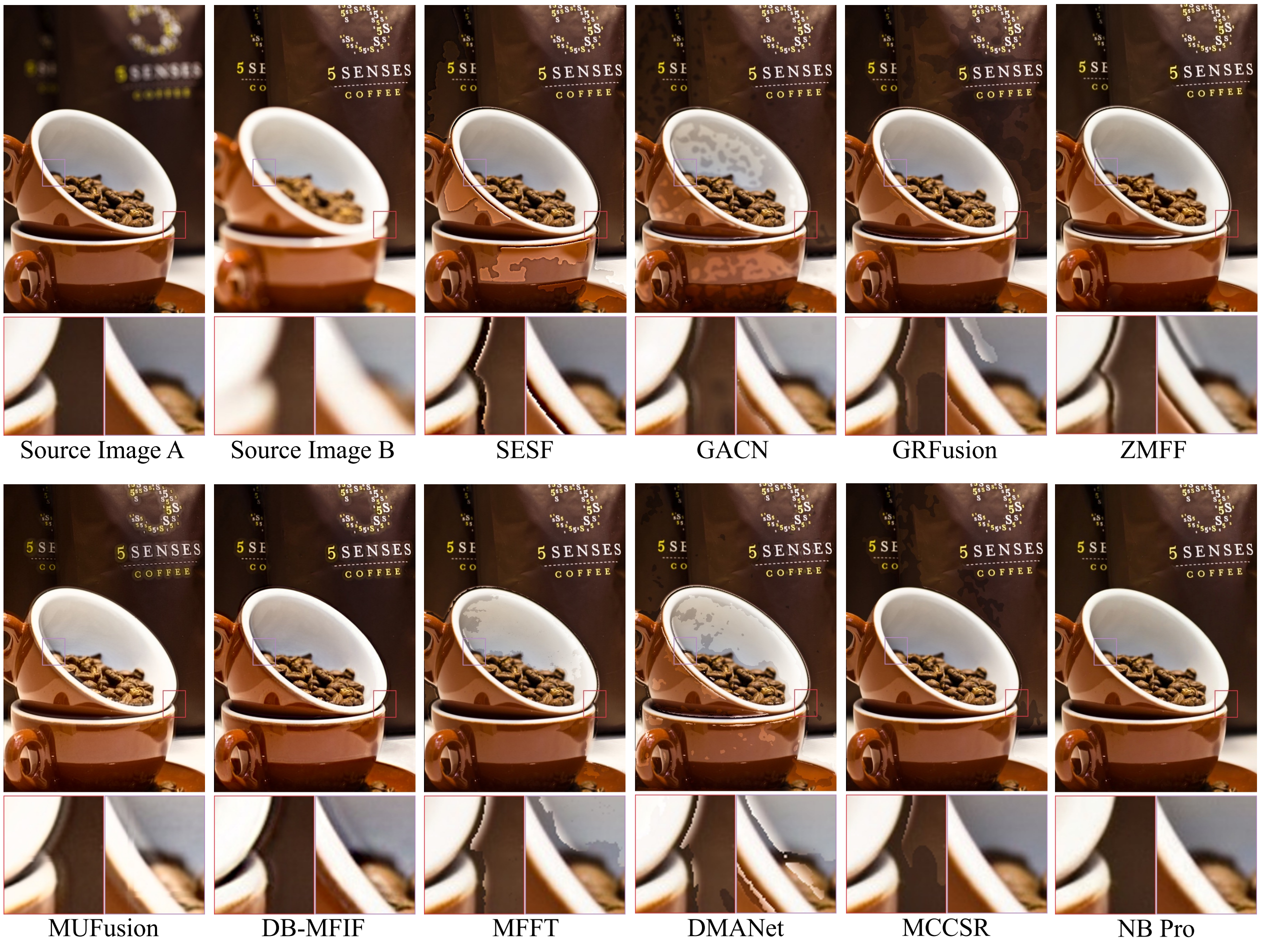


**Figure 34** Visualization of the fusion images of the 'coffee cup' sample from the MFFW dataset. Two enlarged views are shown to reveal critical details.

on source-reference metrics uncovers a critical limitation inherent to zero-shot generative approaches: the prioritization of generative freedom over strict pixel-level adherence to source inputs. While the model can hallucinate plausible high-frequency details (e.g., enhancing lawn textures), it occasionally alters gradients or structures that require preservation.

This quantitative discrepancy stems from the fundamental conflict between the strict consistency required by traditional fusion tasks and the stochastic nature of generative models. First, despite prompt-based constraints, the model's high degree of freedom can lead to semantic alterations in regions that should be preserved. Second, source images are rarely perfect; NB Pro often performs generative enhancement (e.g., super-resolution or denoising) to supplement details. However, traditional reference-based metrics penalize these visual improvements as errors because they deviate from the imperfect source.

Qualitatively, NB Pro demonstrates superior capability in handling complex scenarios, particularly those affected by the Defocus Spread Effect. By effectively mitigating boundary artifacts, dark ghosting, and unnatural transitions common in traditional algorithms, the model showcases a superior semantic understanding of scene structure.

Despite these strengths, instability in focus detection remains a challenge; the model occasionally blurs clear regions, suggesting failures in the attention mechanism's pixel localization. Ultimately, the misalignment between visual superiority and metric penalties suggests that current evaluation frameworks are insufficient for Generative AI. Future work must focus on better constraining the generative process and developing novel metrics capable of distinguishing between hallucinatory errors and generative enhancements.

# 15 Infrared-Visible Image Fusion

## 15.1 Introduction

In the field of modern computer vision, multi-modal image fusion technology plays an increasingly critical role. Infrared-Visible Image Fusion (IVIF) aims to synergize the thermal radiation information from infrared images with the texture details and color information from visible images. Infrared sensors capture thermal signatures of objects and are robust against varying lighting conditions and adverse weather (such as smoke or darkness), effectively highlighting targets. Conversely, visible light sensors provide rich high-frequency details and scene descriptions that align with human visual perception. By fusing these two complementary modalities, the resulting images not only possess all-weather scene perception capabilities but also significantly enhance the accuracy and robustness of target detection, autonomous driving navigation, and security surveillance systems.

Traditional IVIF methods (such as multi-scale transform [98] and sparse representation [80]) often struggle to balance the saliency of thermal targets with the fidelity of background textures, frequently resulting in artificial artifacts. In recent years, deep learning-based methods (such as CNNs [108, 135, 205], GANs [83, 107, 109]) and transformers [84, 142, 158] have achieved performance breakthroughs but still face challenges in cross-modal feature alignment and detail preservation. With the explosion of generative AI technologies, particularly Diffusion Models, image generation quality has reached unprecedented heights. However, existing high-compute models are often bulky and difficult to run in real-time on edge devices. Furthermore, balancing generative quality with physical fidelity in fusion tasks under specific physical constraints remains a pressing problem.

Against this backdrop, Google's newly released Nano Banana Pro has garnered widespread attention in the industry. Nano Banana Pro employs optimized Latent Diffusion technology and efficient attention mechanisms, specifically designed to handle high dynamic range and multi-modal inputs. Its uniqueness lies in its ability to understand semantic context more precisely, suggesting that in image fusion tasks, it may preserve the edge information of infrared thermal sources more effectively than its predecessors while naturally integrating visible textures.

Although Nano Banana Pro has demonstrated impressive performance in general image generation, its effectiveness in the specific scientific task of Infrared-Visible Image Fusion has not yet been systematically verified. This report aims to bridge this gap by comprehensively evaluating the actual performance of Nano Banana Pro in IVIF tasks through qualitative analysis (visual effects) and quantitative assessment. We will focus on examining its fusion quality, noise control capabilities, and inference efficiency across various lighting scenarios to assess its potential as a foundation model for next-generation image fusion.

## 15.2 Quantitative Results

**Table 17** Quantitative comparison on the MSRS, RoadScene, and M$^3$FD datasets. The best results are in **black bold**.

| **Method** | **MSRS** | | | | | | **RoadScene** | | | | | | **M$^3$FD** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $EN$ | $SD$ | $SF$ | $AG$ | $SCD$ | $VIF$ | $EN$ | $SD$ | $SF$ | $AG$ | $SCD$ | $VIF$ | $EN$ | $SD$ | $SF$ | $AG$ | $SCD$ | $VIF$ |
| SDN [189] | 5.25 | 17.35 | 8.67 | 2.67 | 0.99 | 0.50 | 7.30 | 44.06 | 14.58 | 5.80 | 1.37 | 0.61 | 6.87 | 36.22 | 15.32 | 5.61 | 1.41 | 0.55 |
| TarD [95] | 5.28 | 25.22 | 5.98 | 1.83 | 0.71 | 0.42 | 7.26 | 47.44 | 11.11 | 4.14 | 1.40 | 0.56 | 6.80 | 41.77 | 8.65 | 3.17 | 1.35 | 0.51 |
| DeF [92] | 6.46 | 37.63 | 8.60 | 2.80 | 1.35 | 0.77 | 7.36 | 47.03 | 10.99 | 4.38 | 1.62 | 0.63 | 6.90 | 36.81 | 9.85 | 3.65 | 1.42 | 0.58 |
| Meta [204] | 5.65 | 24.97 | 9.99 | 3.40 | 1.14 | 0.31 | 6.88 | 31.97 | 14.38 | 5.57 | 0.92 | 0.55 | 6.73 | 30.56 | 16.48 | 6.02 | 1.31 | 0.65 |
| CDDF [206] | 6.70 | 43.38 | 11.56 | 3.73 | 1.62 | **1.05** | **7.52** | 54.42 | 14.97 | 5.81 | 1.65 | **0.66** | 7.04 | 42.02 | 16.56 | 5.84 | 1.41 | 0.65 |
| LRR [82] | 6.19 | 31.78 | 8.46 | 2.63 | 0.79 | 0.54 | 7.12 | 39.16 | 11.41 | 4.37 | 1.46 | 0.45 | 6.58 | 30.28 | 11.83 | 4.21 | 1.34 | 0.54 |
| MURF [170] | 5.04 | 16.37 | 8.31 | 2.67 | 0.86 | 0.40 | 6.91 | 33.34 | 13.88 | 5.37 | 1.04 | 0.52 | 6.59 | 28.89 | 11.82 | 4.81 | 1.21 | 0.39 |
| DDFM [207] | 6.19 | 29.26 | 7.44 | 2.51 | 1.45 | 0.73 | 7.24 | 42.43 | 10.68 | 4.15 | 1.64 | 0.62 | 6.82 | 32.68 | 10.07 | 3.71 | 1.35 | 0.60 |
| SegM [96] | 5.95 | 37.28 | 11.10 | 3.47 | 1.57 | 0.88 | 7.29 | 46.14 | 14.47 | 5.57 | 1.61 | 0.65 | 6.88 | 36.20 | 16.19 | 5.83 | 1.38 | **0.75** |
| EMMA [208] | 6.71 | 44.13 | 11.56 | 3.76 | **1.63** | 0.97 | **7.52** | 54.81 | 15.21 | 5.83 | **1.69** | **0.66** | **7.12** | **44.01** | **16.92** | **6.23** | **1.48** | 0.66 |
| **NB Pro** | **6.85** | **44.95** | **14.39** | **4.56** | **1.15** | **0.58** | **7.39** | **56.98** | **21.81** | **7.07** | **0.83** | **0.51** | **6.98** | **43.44** | **15.68** | **5.06** | **0.75** | **0.38** |

Following [208], we conduct experiments on three mainstream benchmarks: MSRS [136], RoadScene [169] and M$^3$FD [95] datasets, and leverage six popular metrics for assignment, including non-reference metrics $EN$, $SD$, $SF$, $AG$ and source-reference metrics $SCD$, $VIF$. We compare the fusion results of Nano Banana Pro (NB Pro) with 10 other state-of-the-art and representative IVIF methods, where SDNet and DeFusion

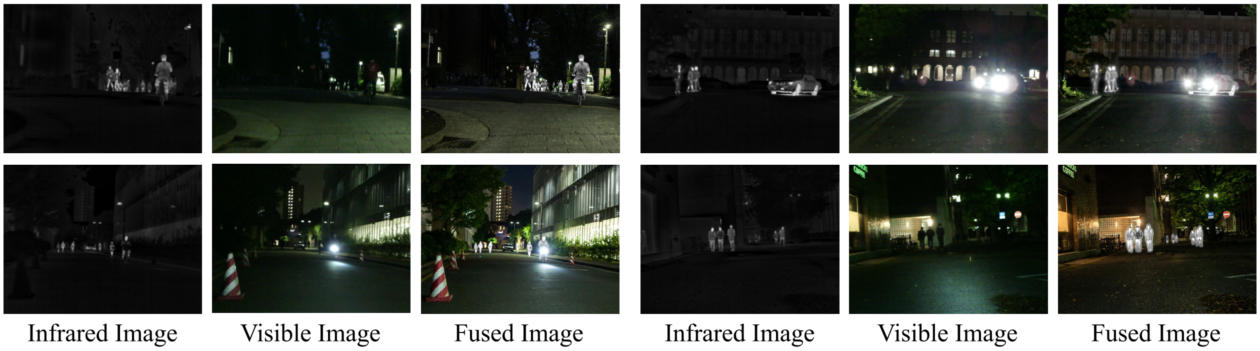


**Figure 35** Visualization examples of image fusion results by Nano Banana Pro in the MSRS dataset.

are CNN-based methods, TarDAL is a GAN-based method, CDDFuse is a CNN-transfomer hybrid method, DDFM is a diffusion-based and training-free method, and the remaining ones are model or task driven approaches.

As shown in Tab. 17, the results demonstrate that NB Pro exhibits overwhelming superiority in non-reference metrics representing image information content and texture details, particularly on the MSRS and RoadScene datasets. On the MSRS dataset, NB Pro secures the top rank in all four of these metrics. Notably, its $EN$ reaches 6.85 and $AG$ hits 4.56, significantly surpassing the runner-up. On the RoadScene dataset, this advantage is even more pronounced. NB Pro achieves an $SF$ score of 21.81, outperforming the nearest competitor (EMMA, 15.21) by nearly 43%. This indicates that the fused images generated by NB Pro possess extremely high clarity and contrast. It is capable of mining and reconstructing rich high-frequency edge information from source images, attributing to the acute capability of its powerful generative architecture in capturing latent features.

However, we also observe an intriguing phenomenon: while NB Pro leads by a wide margin in detail metrics, it scores relatively lower in source-reference metrics, specifically $SCD$ and $VIF$. For instance, on the MSRS dataset, its $VIF$ is only 0.58, and and it drops to 0.38 on M$^3$FD. This reflects the inherent characteristic of generative models: while the model dramatically enhances texture and human-perceived sharpness, this reconstruction process may introduce stylized features or pixel-level deviations not present in the source images, leading to reduced correlation with the original infrared/visible inputs. In contrast, traditional methods, while not as sharp as NB Pro, demonstrate more robustness in maintaining original pixel fidelity.

NB Pro's performance varies across different scenarios. It performs best on MSRS (often containing night-time and complex lighting scenes) and RoadScene, suggesting its proficiency in handling high dynamic range scenes requiring edge enhancement. On the M$^3$FD dataset, although it maintains the second-best score in $SD$(43.44), its overall dominance is less distinct compared to the other two datasets. This implies there may still be room for parameter fine-tuning in specific types of multi-modal target detection scenarios.

## 15.3 Qualitative Results

While NB Pro yields impressive fusion results in most scenarios, certain limitations persist. To intuitively demonstrate the perceptual quality of the fused images, Fig. 35 presents qualitative results on the MSRS dataset. As illustrated in the first row, the method effectively handles extreme lighting conditions. It accurately captures pedestrian targets hidden in low-light regions of the visible spectrum, establishing sharp contours for thermal targets while enhancing overall background illumination. Simultaneously, it restores clear details and textures in over-exposed areas, such as those adjacent to vehicle headlights. However, suboptimal cases are observed in the second row. Although the thermal targets remain highlighted, the first sample exhibits excessive sharpening of the building structure, leading to slight over-exposure. In the second example, a distinct halo effect emerges around the pedestrian, manifesting as unnatural bright fringes surrounding the thermal target.

Fig. 36 further visualize the fusion performance on the other two datasets. In general, NB Pro demonstrates

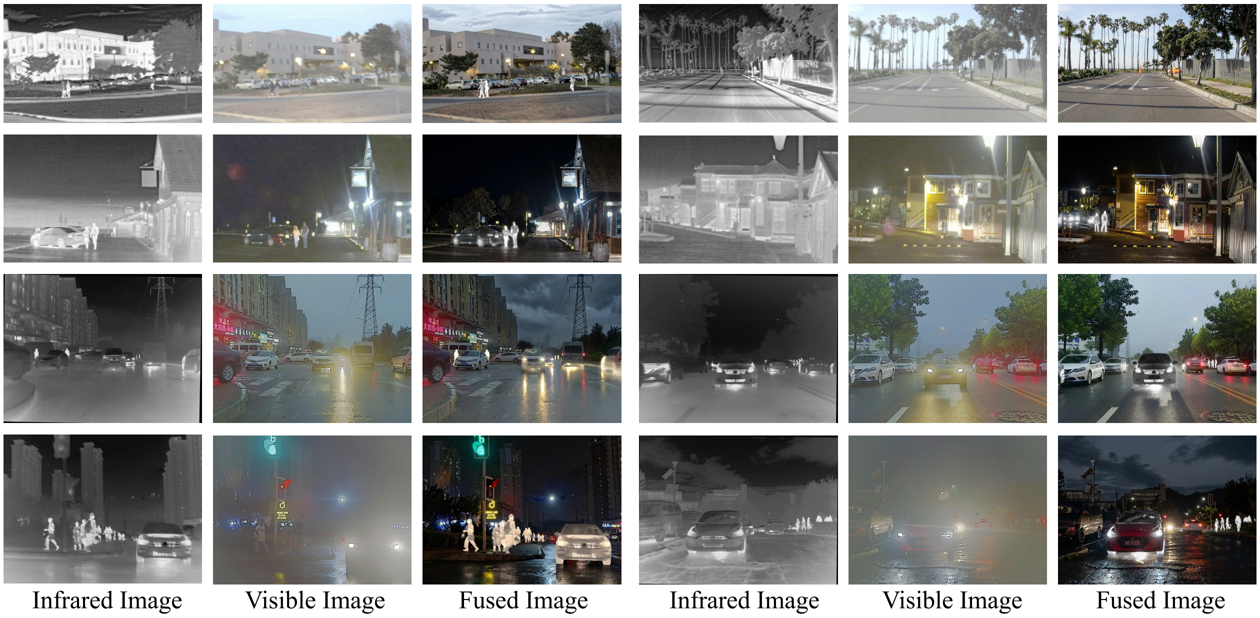


**Figure 36** Visualization examples of image fusion results by Nano Banana Pro in the RoadScene and M$^3$FD datasets.

superior performance in terms of overall visual quality and realism, excelling in infrared target recovery and background texture reconstruction. Nevertheless, deficiencies persist in certain fine-grained details. The model may occasionally fail to faithfully utilize and preserve the source information, leading to the introduction of unnatural hallucinations or artifacts.

## 15.4 Analyses

This report presents a comprehensive evaluation of Google's Nano Banana Pro in Infrared-Visible Image Fusion. The experimental outcomes reveal a pronounced performance dichotomy:

On one hand, leveraging powerful generative priors, NB Pro demonstrates overwhelming superiority in non-reference metrics. It successfully circumvents the bottlenecks of traditional methods regarding night-time enhancement and texture reconstruction, yielding fused images of exceptional contrast and clarity. On the other hand, this aggressive generation strategy incurs a fidelity cost. Lower scores in source consistency metrics, combined with qualitative artifacts such as excessive sharpening and halo effects, indicate that the model sacrifices pixel-level fidelity to the original physical signals in exchange for perceptual appeal.

The performance of NB Pro catalyzes a re-evaluation of the current IVIF landscape:

Traditional methods fundamentally operate as signal processing routines aiming to preserve pixel intensity. However, NB Pro introduces a paradigm of semantic generation. Rather than merely superimposing pixels, it interprets the scene context to re-synthesize the image. This explains its capability to recover astonishing details alongside its propensity for hallucinations. Future research must focus on integrating physical constraints, enforcing strict adherence to thermal distribution laws while exploiting generative capabilities.

While visually striking, NB Pro's outputs raise concerns for safety-critical applications like autonomous driving. Perceptual pleasantness does not equate to operational reliability. Artifacts or over-sharpening can trigger false positives in detection algorithms or obscure small targets. Consequently, evaluation standards must evolve beyond visual quality to include Machine Perception Metrics, directly validating the utility of fused images on downstream tasks.

Our findings highlight the insufficiency of the current evaluation framework. Metrics like SCD and VIF, which penalize pixel-level misalignment, are overly rigid for generative models. As Generative AI becomes prevalent, there is an imperative need for novel No-Reference Image Quality Assessment metrics that prioritize semantic consistency and naturalness over strict pixel-wise alignment.

# 16 Discussion

This comprehensive empirical study, through systematic zero-shot evaluation across 14 diverse low-level vision tasks, elucidates the dual nature of Nano Banana Pro as a generalist generative model: **it excels in perceptual quality but lags significantly in traditional pixel-fidelity metrics.** This core finding not only quantifies the current capabilities of large-scale generative models within the low-level vision domain but also prompts profound reflections on task definitions, evaluation paradigms, and future model evolution.

## 16.1 Generative vs. Regression Paradigms

Our work highlights intrinsic difference between generative and traditional low-level vision models. Traditional methods predominantly follow a regression paradigm. They learn a deterministic mapping from degraded inputs to clean references via pixel-level supervision, aligning their optimization objective directly with metrics like PSNR and SSIM. In contrast, Nano Banana Pro embodies a generative paradigm. Its training core involves learning the joint distribution of large-scale image data and performing conditional synthesis based on semantic priors. Its goal is to produce plausible and visually pleasing images, not to achieve pixel-wise alignment with a specific reference. Consequently, in regions with severe information loss, traditional methods, constrained by input information, often yield blurry or insipid results. The generative model, however, can leverage its robust world knowledge to hallucinate plausible details, leading to subjectively superior outputs that constitute a deviation from the canonical ground truth.

## 16.2 The Potential Misguidance of Traditional Metrics

Our results strongly challenge the universal applicability of full-reference metrics (e.g., PSNR, SSIM), which are predominant in current low-level vision research. These pixel-difference-based metrics carry a strong implicit assumption: the existence of a single, pixel-perfect ground truth. This assumption is problematic for evaluating generative solutions:

**Ground truth is not the unique optimum for generative repair.** For regions with catastrophic information loss, multiple visually plausible and contextually correct reconstructions may exist. A generative model provides one such possibility, yet it is penalized by the metric as incorrect.

**The metrics are misaligned with human perception.** As shown in previous sections, Nano Banana Pro achieves excellent scores on No-Reference perceptual metrics (e.g., NIQE, NIMA), often surpassing specialized models. This indicates its outputs possess superior statistical naturalness and aesthetic appeal. The drop in PSNR can sometimes be attributed solely to the model's reasonable global color adjustment, mild denoising, or detail enhancement, which are improvements that are paradoxically penalized.

**The quality of dataset ground truth itself.** Ground truth images in many real-world datasets contain residual noise, slight blur, or imperfect color balance. A generative model producing a cleaner version constitutes a perceptual enhancement but is scored as a fidelity loss.

Therefore, **judging generative low-level vision models solely by traditional metrics may be both unfair and misleading.** This calls for the community to establish a new generation of evaluation frameworks.

## 16.3 Operational Scope and Limitations of Nano Banana Pro

Our evaluation clearly defines the scope of Nano Banana Pro, shaped by a fundamental compromise: **it favors semantic plausibility and visual appeal over precise pixel-level fidelity.** This positions the model as highly effective for creative and perceptual tasks, such as artistic image enhancement, restoration of severely degraded photos, and scenarios where a visually compelling result is more critical than strict accuracy. Its ability to perform these tasks without specialized training also makes it a practical tool for rapid prototyping.

However, these capabilities come with inherent constraints. The model is not suitable for applications demanding rigorous factual accuracy, including forensic examination, scientific imaging, or any context where the output must correspond exactly to the original scene data. Its generative approach can introduce alterations, such as softened boundaries in super-resolution, altered text in deblurring, or non-physical color shifts, that prioritize visual completeness over authentic reproduction. In essence, **Nano Banana Pro serves as a powerful semantic reconstructor and enhancer for common visual applications, but it is not designed for high-precision tasks where strict fidelity is paramount.**

## 16.4 Future Research Directions

The findings of this study point to several critical directions for future work:

**Exploration of Hybrid Architectures.** The future all-rounder may not be a purely generative model but a generative-regression hybrid. For instance, a lightweight regression network could first recover basic structure and color in the front-end, followed by a conditional generative model for detail enhancement and beautification in the back-end. This process should be constrained by physics-informed loss functions to curb arbitrariness.

**Prompt Engineering and Controllable Generation.** It is important to note that the present evaluation reflects a conservative estimate of the model's capability, as we did not engage in meticulous prompt tuning or employ multi-round inference to cherry-pick optimal outputs. Our fixed, simple prompts represent a pragmatic but unoptimized use case. Future work should therefore systematically explore how carefully designed textual instructions, visual cues, or interactive refinement can more effectively steer the generative process. Enhancing such controllability will be key to reducing unwanted variability in color, structure, and texture—ultimately improving the reliability and practical utility of generative models in restoration-sensitive applications.

**Innovation in Evaluation Frameworks for Generative Models.** The rise of generative models calls for a fundamental shift in how we evaluate their output. Traditional metrics, which rely on a single ground truth, fall short when assessing models that can produce multiple plausible reconstructions from a degraded image. We urgently need new benchmarks that reflect this reality, for instance, datasets that include several expert-approved restoration options for a given input. At the same time, evaluation should move beyond pixel-level fidelity alone. Developing unified metrics that capture both perceptual quality and distortion would provide a more nuanced view of a model's performance across the quality–fidelity spectrum.

## 16.5 Conclusion

The evaluation of Nano Banana Pro signals a paradigm shift: foundational generative models are redefining the boundaries of low-level vision. Functioning less as a traditional restoration tool and more as a semantic reconstruction engine, the model leverages deep generative priors to synthesize visual content rather than merely recovering pixels. This emergence challenges the community to reconsider the fundamental metric of success: is it absolute pixel fidelity, or the maximization of perceptual plausibility?.

Our empirical results confirm that while Nano Banana Pro trails domain-specific experts in zero-shot pixel fidelity, it demonstrates exceptional potential in perceptual quality, particularly when handling extreme degradations and cross-task generalization. Consequently, the trajectory of the field lies not in a binary choice between paradigms, but in strategic integration. The next generation of robust vision systems must bridge the semantic imagination of generative models with the physical constraints and precision of specialized networks.

Ultimately, Nano Banana Pro is a double-edged sword. It has successfully raised the ceiling of perceptual quality for complex visual tasks, yet it has not secured the floor of stability required for forensic precision.

# 17 Appendix

## 17.1 Prompts for Each Task

**Dehazing:** *"This is an image with hazy, which reduces clarity and contrast. Please completely remove this atmospheric haze from the image while meticulously preserving all other elements without any alteration. The post-processing must be strictly limited to haze removal only. Ensure that every other aspect of the image remains entirely unchanged, including but not limited to the original composition, all subjects and objects within the scene. The final output should be a clear, natural-looking version of the original image with its core content perfectly intact."*

**Super-Resolution:** *"Please upscale this low-resolution image to a high-quality 1k (1024x1024) resolution. Perform super-resolution processing to significantly enhance clarity, remove pixelation, and sharpen textures, while strictly preserving the original content, composition, and colors unchanged. Do not alter the subject's features or hallucinate new elements."*

**Deraining:** *"This is a rainy image. Please remove the rain streaks and raindrops while keeping all other elements, the original color tone, lighting, and atmosphere unchanged."*

**Shadow Removal:** *"This is an image with prominent cast shadows. Please remove these shadows from the image while ensuring all other elements in the scene remain completely unchanged and consistent. The core requirement is to eliminate the shadow effects while preserving the inherent properties of all objects and the background. The final output should be a clean, evenly lit version of the original image, without any dark occlusions or shaded areas."*

**Motion Deblurring:** *"This is an image with blurring and lack of clarity due to motion. Transform this image into a sharp, static photograph. Please carefully observe this picture to eliminate motion streaks and recover details in any blurred areas—regardless of the motion source—while keeping originally sharp elements consistent. CRITICAL EXPOSURE LOCK: Strictly prohibit any form of exposure correction, HDR tone mapping, or lighting enhancement. Do not attempt to recover details in blown-out highlights or brighten dark/underexposed shadows. You must preserve the exact luminance levels of the original image. Zero changes to brightness, contrast, or exposure are allowed."*

**Defocus Deblurring:** *"This is an image with partial blurring and lack of clarity due to camera defocus. Transform this image into an all-in-focus photograph. Please carefully observe this picture to enhance the sharpness of the blurred areas while keeping all other elements consistent. Maintain the original color palette, lighting, and exposure faithfully. No color shifts or brightness changes."*

**Denoising:** *"This is a noisy image, please remove the noise in this image while keep other elements in this image unchanged."*

**Reflection Removal:** *"Strictly remove only the glass reflections and glare overlaying the scene. Do not alter the underlying scene composition, object details, geometry, or color grading in any way. The objective is to make the glass invisible while maintaining pixel-perfect fidelity to the original background. Zero tolerance for hallucinations, artistic interpretation, or style changes. Retain the exact original perspective and lighting conditions of the scene behind the glass. Treat this as a forensic image restoration."*

**Flare Removal:** *"Identify and remove lens flare and glare artifacts generated by multiple distinct light sources throughout the entire image. For every single light source causing flare, eliminate the optical interference (such as ghosting, halos, and streaks) while strictly maintaining the natural illumination and intensity of each individual light. Seamlessly restore the obscured background textures behind all flare instances to ensure global consistency."*

**Low Light Image Enhancement:** *"This is a low-light image, please turn this image into a normal image while keeping other elements unchanged."*

**Underwater Image Enhancement:** *"This is an underwater image with obvious color cast, low contrast, and scattered fog. Please carefully analyze the main scene, eliminate the interference caused by water absorption and suspended particles, and restore a clear, fog-free version with true colors. Ensure that all elements except for the degradation factors are consistent with the real underwater environment, without introducing new* *artifacts or over-enhancement.CRITICAL ELEMENT LOCK: remove only the color shift and fog; do not add, remove, or alter any original object, edge, texture."*

**HDR Imaging:** *"This is an image suffering from limited dynamic range, with lost details in highlights and shadows. Transform this image into a high dynamic range photograph. Please carefully analyze this picture to expand its dynamic range, recover details in both clipped highlights and underexposed shadows, while preserving the integrity of properly exposed areas.CRITICAL CONTENT LOCK:Strictly prohibit any form of scene content alteration, object addition, or deletion. Do not change the shape, position, texture, or color relationships of any original elements. Forbid the introduction of any artistic styles, filter effects, or non-physical halos.CRITICAL PROCESSING PRINCIPLES:1. Highlight Recovery: Intelligently reconstruct plausible texture and detail lost in overexposed areas (e.g., sky, windows, light sources), but do not alter their fundamental form or color.2. Shadow Enhancement: Selectively lift the brightness of shadow areas to reveal concealed details, while must retaining the depth and original distribution of the shadows.3. Midtone Preservation: Maintain the natural contrast and color of midtone areas. Avoid introducing an overall flat/gray look or unnatural local contrast.FINAL OBJECTIVE: Produce a detail-rich, balanced, and natural-looking image that appears as if the same scene was captured in a single shot with professional HDR equipment, not artificially composited."*

**Multi-Focus Image Fusion:** *"Act as an advanced Computer Vision expert specialized in Multi-Focus Image Fusion(MFIF). Your task is to process a pair of source images with different focal depths (e.g., near-focus and far-focus) and merge them into a single, high-quality 'all-in-focus' image. You must analyze the local sharpness and high-frequency details of each input to accurately identify the clearest regions, constructing a precise decision map that selects the best pixels from the respective sources. Ensure to apply boundary refinement techniques to guarantee smooth transitions between fused areas. The final output must be a seamless, fully focused image that strictly preserves the original color and structural fidelity while being completely free of visual artifacts such as ghosting, halos, or unnatural stitching seams."*

**Infrared-Visible Image Fusion:** *"Act as an expert in Infrared and Visible Image Fusion (IVIF). Your task is to generate a single high-quality fused image based on the provided Infrared (IR) and Visible (VIS) source images. The core objective is to integrate complementary features: you must preserve the high-intensity thermal saliency (such as pedestrians and vehicles) from the IR image to ensure target detectability, while simultaneously injecting the high-frequency textural details (such as edges, vegetation, and building structures) from the VIS image to ensure background clarity. The fusion process must balance the intensity distribution to look perceptually natural, strictly minimizing artifacts like ghosting, halos, or noise. The final output should be a sharp, noise-free image that combines the high contrast of the thermal targets with the rich structural details of the visible scene."*

## 17.2 Contributors

**Jialong Zuo:** Project Leader. **Haoyou Deng:** Document Polish.

**Hanyu Zhou:** Denoising and Low Light Enhancement. **Jiaxin Zhu:** Motion Deblurring and Defocus Deblurring.

**Yicheng Zhang:** Multi-focus Image Fusion and Infrared-Visible Image Fusion.

**Yiwei Zhang:** Underwater Image Enhancement and HDR Imaging.

**Yongxin Yan:** Dehazing and Shadow Removal. **Kaixing Huang:** Super Resolution.

**Weisen Chen:** Deraining. **Yongtai Deng:** Reflection Removal. **Rui Jin:** Flare Removal.

**Nong Sang:** Advisor. **Changxin Gao:** Advisor.

## References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1692–1700, 2018.

[2] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *European conference on computer vision*, pages 111–126. Springer, 2020.

[3] Abdullah Abuolaim, Mahmoud Afifi, and Michael S Brown. Improving single-image defocus deblurring: How dual-pixel images help through multi-task learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1231–1239, 2022.

[4] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 126–135, 2017.

[5] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert. Enhancing underwater images and videos by fusion. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 81–88, 2012.

[6] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert. Color balance and fusion for underwater image enhancement. *IEEE Trans. Image Process.*, 27(1):379–393, 2018.

[7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[8] Dana Berman, Shai Avidan, et al. Non-local image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1674–1682, 2016.

[9] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 6228–6237, 2018.

[10] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Fredo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *CVPR 2011*, pages 97–104, 2011. doi: 10.1109/CVPR.2011.5995332.

[11] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018.

[12] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3086–3095, 2019.

[13] Jie Cai, Kangning Yang, Ling Ouyang, Lan Fu, Jiaming Ding, Huiming Sun, Chiu Man Ho, and Zibo Meng. F2t2-hit: A u-shaped fft transformer and hierarchical transformer for reflection removal. *arXiv preprint arXiv:2506.05489*, 2025.

[14] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12504–12513, 2023.

[15] Haoyu Chen, Jinjin Gu, Yihao Liu, Salma Abdel Magid, Chao Dong, Qiong Wang, Hanspeter Pfister, and Lei Zhu. Masked image training for generalizable deep image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1703, 2023.

[16] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5896–5905, 2023.

[17] Xiang Chen, Jinshan Pan, and Jiangxin Dong. Bidirectional multi-scale implicit neural representations for image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[18] Yin Chen and Rick S Blum. A new automated quality assessment algorithm for image fusion. *Image and vision computing*, 27(10):1421–1432, 2009.

[19] Zeyuan Chen, Yangchao Wang, Yang Yang, and Dong Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7180–7189, 2021.

[20] Zheng Chen, Yulun Zhang, Ding Liu, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Hierarchical integration diffusion model for realistic image deblurring. *Advances in neural information processing systems*, 36:29114–29125, 2023.

[21] Chunyang Cheng, Tianyang Xu, and Xiao-Jun Wu. Mufusion: A general unsupervised image fusion network based on memory unit. *Information Fusion*, 92:80–92, 2023.

[22] R. Cong, W. Yang, W. Zhang, C. Li, C.-L. Guo, Q. Huang, and S. Kwong. Pugan: Physical model-guided underwater image enhancement using gan with dual-discriminators. *IEEE Transactions on Image Processing*, 32: 4472–4485, 2023.

[23] Guangmang Cui, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015.

[24] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10680–10687, 2020.

[25] Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Yihang Luo, and Chen Change Loy. Flare7k++: Mixing synthetic and real datasets for nighttime flare removal and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7041–7055, 2024. doi: 10.1109/TPAMI.2024.3406821.

[26] Yuekun Dai, Dafeng Zhang, Xiaoming Li, Zongsheng Yue, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Peiqing Yang, Zhezhu Jin, Guanqun Liu, and Chen Change Loy. Mipi 2024 challenge on nighttime flare removal: Methods and results. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1144–1152, 2024. doi: 10.1109/CVPRW63382.2024.00121.

[27] Haoyou Deng, Lida Li, Feng Zhang, Zhiqiang Li, Bin Xu, Qingbo Lu, Changxin Gao, and Nong Sang. Towards blind flare removal using knowledge-driven flare-level estimator. *IEEE Transactions on Image Processing*, 2024.

[28] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. *European conference on computer vision (ECCV)*, 2014.

[29] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2157–2167, 2020.

[30] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Location-aware single image reflection removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5017–5026, 2021.

[31] P. L. Drews, E. R. Nascimento, S. S. Botelho, and M. F. M. Campos. Underwater depth estimation and image restoration based on single images. *IEEE Computer Graphics and Applications*, 36(2):24–35, 2016.

[32] Chengyu Fang, Chunming He, Fengyang Xiao, Yulun Zhang, Longxiang Tang, Yuelin Zhang, Kai Li, and Xiu Li. Real-world image dehazing with coherence-based pseudo labeling and cooperative unfolding network. *Advances in Neural Information Processing Systems*, 37:97859–97883, 2024.

[33] Raanan Fattal. Dehazing using color-lines. *ACM transactions on graphics (TOG)*, 34(1):1–14, 2014.

[34] Rich Franzen. *Kodak lossless true color image suite*, volume 5. https://r0k.us/graphics/kodak/, 1999.

[35] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3855–3863, 2017.

[36] Xueyang Fu, Qi Qi, Zheng-Jun Zha, Yurui Zhu, and Xinghao Ding. Rain streak removal via dual graph convolutional network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1352–1360, 2021.

[37] Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22252–22261, 2023.

[38] A. Galdran, D. Pardo, A. Picon, and A. Alvarez-Gila. Automatic red-channel underwater image restoration. *Journal of Visual Communication and Image Representation*, 26:132–145, 2015.

[39] S.-B. Gao, M. Zhang, Q. Zhao, X.-S. Zhang, and Y.-J. Li. Underwater image enhancement using adaptive retinal mechanisms. *IEEE Trans. Image Process.*, 28(11):5580–5595, 2019.

[40] A. S. A. Ghani and N. A. M. Isa. Underwater image quality enhancement through integrated color model with rayleigh distribution. *Applied Soft Computing*, 27:219–230, 2015.

[41] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

[42] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems (NeurIPS)*, 2014.

[43] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5812–5820, 2022.

[44] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1780–1789, 2020.

[45] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps shadow removal. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 710–718, 2023.

[46] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14049–14058, 2023.

[47] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2956–2967, 2012.

[48] Xiaojie Guo and Qiming Hu. Low-light image enhancement via breaking down the darkness. *International Journal of Computer Vision*, 131(1):48–66, 2023.

[49] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. In *European Conference on Computer Vision*, pages 679–695. Springer, 2020.

[50] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

[51] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[52] Yuchen Hong, Haofeng Zhong, Shuchen Weng, Jinxiu Liang, and Boxin Shi. L-differ: Single image reflection removal with language-based diffusion model. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024.

[53] Mohammed Hossny, Saeid Nahavandi, and Douglas Creighton. Comments on 'information measure for performance of image fusion'. *Electronics letters*, 44(18):1066–1067, 2008.

[54] Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *Advances in Neural Information Processing Systems*, 36:79734–79747, 2023.

[55] Jichen Hu, Chen Yang, Zanwei Zhou, Jiemin Fang, Xiaokang Yang, Qi Tian, and Wei Shen. Dereflection any image with diffusion priors and diversified data. *arXiv preprint arXiv:2503.17347*, 2025.

[56] Qiming Hu and Xiaojie Guo. Trash or treasure? an interactive dual-stream strategy for single image reflection separation. *Advances in Neural Information Processing Systems*, 34:24683–24694, 2021.

[57] Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13138–13147, 2023.

[58] Qiming Hu, Hainuo Wang, and Xiaojie Guo. Single image reflection separation via dual-stream interactive transformers. *Advances in Neural Information Processing Systems*, 37:55228–55248, 2024.

[59] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7454–7462, 2018.

[60] Xingyu Hu, Junjun Jiang, Xianming Liu, and Jiayi Ma. Zmff: Zero-shot multi-focus image fusion. *Information Fusion*, 92:127–138, 2023.

[61] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015.

[62] Yihang Huang, Yuanfei Huang, Junhui Lin, and Hua Huang. Deflaremamba: Hierarchical vision mamba for contextually consistent lens flare removal. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 8028–8037, 2025.

[63] Yue Huang, Zi'ang Li, Tianle Hu, Jie Wen, Guanbin Li, Jinglin Zhang, Guoxu Zhou, and Xiaozhao Fang. Single image reflection removal via inter-layer complementarity. *arXiv preprint arXiv:2505.12641*, 2025.

[64] Matthias Hullin, Elmar Eisemann, Hans-Peter Seidel, and Sungkil Lee. Physically-based real-time lens flare rendering. *ACM Trans. Graph.*, 30(4), 2011.

[65] K. Iqbal, M. Odetayo, A. James, R. A. Salam, and A. Z. H. Talib. Enhancing the low quality images using unsupervised colour correction method. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 1703–1709, 2010.

[66] Bernd Jähne. *Digital image processing*. Springer, 2005.

[67] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8346–8355, 2020.

[68] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021.

[69] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021.

[70] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018.

[71] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8878–8887, 2019.

[72] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Alykhan Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4681–4690, 2017.

[73] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2034–2042, 2021.

[74] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE transactions on image processing*, 28(1):492–505, 2018.

[75] C. Li, J. Guo, B. Wang, R. Cong, Y. Zhang, and J. Wang. Single underwater image enhancement based on color cast removal and visibility restoration. *Journal of Electronic Imaging*, 25(3):033012, 2016.

[76] C.-Y. Li, J.-C. Guo, R.-M. Cong, Y.-W. Pang, and B. Wang. Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior. *IEEE Transactions on Image Processing*, 25(12): 5664–5677, 2016.

[77] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E Hopcroft. Single image reflection removal through cascaded refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3565–3574, 2020.

[78] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, nov 2019. doi: 10.1109/TIP.2019.2955241.

[79] H. Li, J. Li, and W. Wang. A Fusion Adversarial Underwater Image Enhancement Network with a Public Test Dataset. *arXiv preprint*, 2019. doi: 10.48550/arXiv.1906.06819. arXiv:1906.06819.

[80] Huafeng Li, Yitang Wang, Zhao Yang, Ruxin Wang, Xiang Li, and Dapeng Tao. Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion. *IEEE Transactions on Instrumentation and Measurement*, 69(4):1082–1102, 2019.

[81] Huafeng Li, Dan Wang, Yuxin Huang, Yafei Zhang, and Zhengtao Yu. Generation and recombination for multifocus image fusion with free number of inputs. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6009–6023, 2023.

[82] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):11040–11052, 2023.

[83] Jing Li, Hongtao Huo, Chang Li, Renhua Wang, and Qi Feng. Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Transactions on Multimedia*, 23:1383–1396, 2020.

[84] Jing Li, Jianming Zhu, Chang Li, Xun Chen, and Bin Yang. Cgtf: Convolution-guided transformer for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2022.

[85] Jinxing Li, Xiaobao Guo, Guangming Lu, Bob Zhang, Yong Xu, Feng Wu, and David Zhang. Drpl: Deep regression pair learning for multi-focus image fusion. *IEEE Transactions on Image Processing*, 29:4816–4831, 2020.

[86] Mining Li, Ronghao Pei, Tianyou Zheng, Yang Zhang, and Weiwei Fu. Fusiondiff: Multi-focus image fusion using denoising diffusion probabilistic models. *Expert Systems with Applications*, 238:121664, 2024.

[87] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 254–269, 2018.

[88] Xiaoyu Li, Bo Zhang, Jing Liao, and Pedro V. Sander. Let's see clearly: Contaminant artifact removal for moving cameras. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1991–2000, 2021. doi: 10.1109/ICCV48922.2021.00202.

[89] Xin Li, Bingchen Li, Xin Jin, Cuiling Lan, and Zhibo Chen. Learning distortion invariant representation for image restoration from a causality perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1714–1724, 2023.

[90] Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2736–2744, 2016.

[91] Yu Li, Ming Liu, Yaling Yi, Qince Li, Dongwei Ren, and Wangmeng Zuo. Two-stage single image reflection removal with reflection-aware guidance. *Applied Intelligence*, 53(16):19433–19448, 2023.

[92] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *European conference on computer vision*, pages 719–735. Springer, 2022.

[93] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. In *arXiv preprint arXiv:2308.15070*, 2023.

[94] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[95] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5811, 2022.

[96] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8115–8124, 2023.

[97] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10561–10570, 2021.

[98] Yu Liu, Shuping Liu, and Zengfu Wang. A general framework for image fusion based on multi-scale transform and sparse representation. *Information fusion*, 24:147–164, 2015.

[99] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, 2017.

[100] Yanzuo Lu, Xin Xia, Manlin Zhang, Huafeng Kuang, Jianbin Zheng, Yuxi Ren, and Xuefeng Xiao. Hyper-bagel: A unified acceleration framework for multimodal understanding and generation. *arXiv preprint arXiv:2509.18824*, 2025.

[101] Zhengyang Lu, Weifan Wang, Tianhao Guo, and Feng Wang. Single-image reflection removal via self-supervised diffusion models. *The Journal of Supercomputing*, 81(1):338, 2025.

[102] Simin Luan, Cong Yang, Zeyd Boukhers, Xue Qin, Dongfeng Cheng, Wei Sui, and Zhijun Li. Gyroscope-assisted motion deblurring network. *CoRR*, 2024.

[103] Yu Luo, Yong Xu, and Hui Ji. Removing rain from a single image via discriminative sparse coding. In *Proceedings of the IEEE international conference on computer vision*, pages 3397–3405, 2015.

[104] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. In *Proceedings of the 40th International Conference on Machine Learning*, pages 23045–23066, 2023.

[105] Boyuan Ma, Yu Zhu, Xiang Yin, Xiaojuan Ban, Haiyou Huang, and Michele Mukeshimana. Sesf-fuse: An unsupervised deep model for multi-focus image fusion. *Neural Computing and Applications*, 33(11):5793–5804, 2021.

[106] Boyuan Ma, Xiang Yin, Di Wu, Haokai Shen, Xiaojuan Ban, and Yu Wang. End-to-end learning for simultaneously generating decision map and multi-focus image fusion result. *Neurocomputing*, 470:204–216, 2022.

[107] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information fusion*, 48:11–26, 2019.

[108] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020.

[109] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29: 4980–4995, 2020.

[110] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9 (7):1200–1217, 2022.

[111] Armin Mehri, Parichehr B Ardakani, and Angel D Sappa. Mprnet: Multi-path residual network for lightweight image super resolution. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2704–2713, 2021.

[112] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a completely blind image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[113] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12826–12835, 2020.

[114] Vineeth Murali and PV Sudeep. Image denoising using dncnn: An exploration study. In *Advances in Communication Systems and Networks: Select Proceedings of ComNet 2019*, pages 847–859. Springer, 2020.

[115] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017.

[116] Mansour Nejati, Shadrokh Samavi, and Shahram Shirani. Multi-focus image fusion using dictionary-based sparse representation. *Information fusion*, 25:72–84, 2015.

[117] K. Panetta, C. Gao, and S. Agaian. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, 41(3):541–551, 2015. doi: 10.1109/JOE.2015.2410644.

[118] L. Peng, C. Zhu, and L. Bian. U-shape transformer for underwater image enhancement. *IEEE Transactions on Image Processing*, 32:3066–3079, 2023. doi: 10.1109/TIP.2023.3276332.

[119] Yuwei Qiu, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, and Zhi Jin. Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12802–12813, 2023.

[120] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4067–4075, 2017.

[121] Yuhui Quan, Zicong Wu, and Hui Ji. Gaussian kernel mixture network for single image defocus deblurring. *Advances in Neural Information Processing Systems*, 34:20812–20824, 2021.

[122] Yuhui Quan, Zicong Wu, and Hui Ji. Neumann network with recursive kernels for single image defocus deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5754–5763, 2023.

[123] Yuhui Quan, Xin Yao, and Hui Ji. Single image defocus deblurring via implicit neural inverse kernels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12600–12610, 2023.

[124] Yuhui Quan, Zicong Wu, Ruotao Xu, and Hui Ji. Deep single image defocus deblurring via gaussian kernel mixture learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[125] Yuhui Quan, Xi Wan, Zitao Tang, Jinxiu Liang, and Hui Ji. Multi-focus image fusion via explicit defocus blur modelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6657–6665, 2025.

[126] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3937–3946, 2019.

[127] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European conference on computer vision*, pages 184–201. Springer, 2020.

[128] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[129] Lingyan Ruan, Bin Chen, Jizhou Li, and Miu-Ling Lam. Aifnet: All-in-focus image restoration network using a light field-based dataset. *IEEE Transactions on Computational Imaging*, 7:675–688, 2021.

[130] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16304–16313, 2022.

[131] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang. Domain adaptation for image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2808–2817, 2020.

[132] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5572–5581, 2019.

[133] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2642–2650, 2021.

[134] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023.

[135] Hao Tang, Chengcheng Yuan, Zechao Li, and Jinhui Tang. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition*, 130:108792, 2022.

[136] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022.

[137] Yi Tang, Hiroshi Kawasaki, and Takafumi Iwaguchi. Underwater image enhancement by transformer-based diffusion model with non-uniform sampling for skip strategy. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, pages 5419–5427, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3581783.3612475.

[138] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[139] Chun-Chieh Tsai. Standard images for multifocus image fusion, 2025.

[140] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European conference on computer vision*, pages 146–162. Springer, 2022.

[141] Yael Vinker, Inbar Huberman-Spiegelglas, and Raanan Fattal. Unpaired learning for high dynamic range image tone mapping. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14637–14646, 2021. doi: 10.1109/ICCV48922.2021.01439.

[142] Vibashan Vs, Jeya Maria Jose Valanarasu, Poojan Oza, and Vishal M Patel. Image fusion transformer. In *2022 IEEE International conference on image processing (ICIP)*, pages 3566–3570. IEEE, 2022.

[143] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE international conference on computer vision*, pages 3922–3930, 2017.

[144] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. A model-driven deep neural network for single image rain removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3103–3112, 2020.

[145] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563, 2023.

[146] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision (IJCV)*, pages 1–21, 2024.

[147] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6849–6857, 2019.

[148] Tao Wang, Yong Li, Jingyang Peng, Yipeng Ma, Xian Wang, Fenglong Song, and Youliang Yan. Real-time image enhancer via learnable spatial-aware 3d lookup tables. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2471–2480, 2021.

[149] Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2654–2662, 2023.

[150] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson W.H. Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[151] Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26057–26066, 2024.

[152] 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 (ICCV) Workshops*, pages 1905–1914, 2021.

[153] Yudong Wang, Jichang Guo, Huan Gao, and Huihui Yue. Uiec$^2$-net: Cnn-based underwater image enhancement using two color space. *Signal Processing: Image Communication*, 96:116250, 2021. ISSN 0923-5965.

[154] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2604–2612, 2022.

[155] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25796–25805, 2024.

[156] Zhaohan Wang, Chengjun Chen, and Chenggang Dai. Zero-shot realistic image deblurring with consistency model. *Complex & Intelligent Systems*, 12(1):29, 2026.

[157] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022.

[158] Zhishe Wang, Yanlin Chen, Wenyu Shao, Hui Li, and Lei Zhang. Swinfuse: A residual swin transformer fusion network for infrared and visible images. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022.

[159] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[160] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[161] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2019.

[162] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 101–117, 2020.

[163] Rui-Qi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, and Chongyi Li. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22282–22291, 2023.

[164] Yicheng Wu, Qiurui He, Tianfan Xue, Rahul Garg, Jiawen Chen, Ashok Veeraraghavan, and Jonathan T. Barron. How to train neural networks for flare removal. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2219–2227, 2021. doi: 10.1109/ICCV48922.2021.00224.

[165] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13095–13105, 2023.

[166] Bin Xiao, Haifeng Wu, and Xiuli Bi. Dtmnet: A discrete tchebichef moments-based deep neural network for multi-focus image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 43–51, 2021.

[167] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12978–12995, 2022.

[168] Jie Xiao, Xueyang Fu, Yurui Zhu, Dong Li, Jie Huang, Kai Zhu, and Zheng-Jun Zha. Homoformer: Homogenized transformer for image shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25617–25626, 2024.

[169] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12484–12491, 2020.

[170] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):12148–12166, 2023.

[171] Jun Xu, Hui Li, Zhetong Liang, David Zhang, and Lei Zhang. Real-world noisy image denoising: A new benchmark. *arXiv preprint arXiv:1804.02603*, 2018.

[172] Shuang Xu, Xiaoli Wei, Chunxia Zhang, Junmin Liu, and Jiangshe Zhang. Mffw: A new dataset for multi-focus image fusion. *arXiv preprint arXiv:2002.04780*, 2020.

[173] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17714–17724, 2022.

[174] Hanshu Yan, Jingfeng Zhang, Jiashi Feng, Masashi Sugiyama, and Vincent YF Tan. Towards adversarially robust deep image denoising. *arXiv preprint arXiv:2201.04397*, 2022.

[175] Cui Yang, Jian-Qi Zhang, Xiao-Rui Wang, and Xin Liu. A novel similarity based quality metric for image fusion. *Information Fusion*, 9(2):156–160, 2008.

[176] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proceedings of the european conference on computer vision (ECCV)*, pages 654–669, 2018.

[177] M. Yang and A. Sowmya. An underwater color image quality evaluation metric. *IEEE Transactions on Image Processing*, 24(12):6062–6071, 2015. doi: 10.1109/TIP.2015.2480136.

[178] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1357–1366, 2017.

[179] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021.

[180] Yang Yang, Chaoyue Wang, Risheng Liu, Lin Zhang, Xiaojie Guo, and Dacheng Tao. Self-augmented unpaired image dehazing via density and depth decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2037–2046, 2022.

[181] Qiaosi Yi, Juncheng Li, Qinyan Dai, Faming Fang, Guixu Zhang, and Tieyong Zeng. Structure-preserving deraining with residue channel prior guidance. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4238–4247, 2021.

[182] Daniyar Zakarin, Thiemo Wandel, Anton Obukhov, and Dengxin Dai. Reflection removal through efficient adaptation of diffusion transformers. *arXiv preprint arXiv:2512.05000*, 2025.

[183] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021.

[184] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[185] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2058–2073, 2020.

[186] Hao Zhai, Wenyi Zheng, Yuncan Ouyang, Xin Pan, and Wanli Zhang. Multi-focus image fusion via interactive transformer and asymmetric soft sharing. *Engineering Applications of Artificial Intelligence*, 133:107967, 2024.

[187] Feng Zhang, Haoyou Deng, Zhiqiang Li, Lida Li, Bin Xu, Qingbo Lu, Zisheng Cao, Minchen Wei, Changxin Gao, Nong Sang, et al. High-resolution photo enhancement in real-time: A laplacian pyramid network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[188] Fengyi Zhang, Hui Zeng, Tianjun Zhang, and Lin Zhang. Clut-net: Learning adaptively compressed representations of 3dluts for lightweight image enhancement. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6493–6501. ACM, 2022.

[189] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10):2761–2785, 2021.

[190] Hao Zhang, Zhuliang Le, Zhenfeng Shao, Han Xu, and Jiayi Ma. Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Information Fusion*, 66:40–53, 2021.

[191] Juncheng Zhang, Qingmin Liao, Haoyu Ma, Jing-Hao Xue, Wenming Yang, and Shaojun Liu. Exploit the best of both end-to-end and map-based methods for multi-focus image fusion. *IEEE Transactions on Multimedia*, 26: 6411–6423, 2024.

[192] 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.

[193] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4791–4800, 2021.

[194] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2737–2746, 2020.

[195] Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Electronic imaging*, 20(2):023016–023016, 2011.

[196] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 586–595, 2018.

[197] W. Zhang, Y. Wang, and C. Li. Underwater image enhancement by attenuated color channel correction and detail preserved contrast enhancement. *IEEE Journal of Oceanic Engineering*, pages 1–18, 2022.

[198] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4786–4794, 2018.

[199] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1632–1640, 2019.

[200] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54:99–118, 2020.

[201] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[202] C. Zhao, W. Cai, C. Dong, and C. Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8281–8291, Seattle, WA, USA, jun 2024. IEEE/CVF. doi: 10.1109/CVPR52729.2024.00813.

[203] Hao Zhao, Mingjia Li, Qiming Hu, and Xiaojie Guo. Reversible decoupling network for single image reflection removal. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26430–26439, 2025.

[204] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13955–13965, 2023.

[205] Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1186–1196, 2021.

[206] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5906–5916, 2023.

[207] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8082–8093, 2023.

[208] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25912–25921, 2024.

[209] Kecheng Zheng, Juan Cheng, and Yu Liu. Unfolding coupled convolutional sparse representation for multi-focus image fusion. *Information Fusion*, 118:102974, 2025.

[210] Yufeng Zheng, Edward A Essock, Bruce C Hansen, and Andrew M Haun. A new metric based on extended spatial frequency and its application to dwt based fusion algorithms. *Information Fusion*, 8(2):177–192, 2007.

[211] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5627–5636, 2022.

[212] Yurui Zhu, Xueyang Fu, Peng-Tao Jiang, Hao Zhang, Qibin Sun, Jinwei Chen, Zheng-Jun Zha, and Bo Li. Revisiting single image reflection removal in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25468–25478, 2024.