

PASDiff: Physics-Aware Semantic Guidance for Joint Real-World Low-Light Face Enhancement and Restoration

Supplementary Material

Yilin Ni^{1,*}, Wenjie Li^{2,*}, Zhengxue Wang³, Juncheng Li⁴,
Guangwei Gao^{3,†}, and Jian Yang³

¹ College of Automation, Nanjing University of Posts and Telecommunications,
Nanjing, China

² School of Artificial Intelligence, Beijing University of Posts and
Telecommunications, Beijing, China

³ PCA Lab, School of Computer Science and Engineering, Nanjing University of
Science and Technology, Nanjing, China

⁴ School of Computer Science and Technology, East China Normal University,
Shanghai, China

{nixiaolin26,lewj2408}@gmail.com, gwgao@njjust.edu.cn

In this supplementary material, we provide additional implementation details, further analysis, and comprehensive visual results to complement the main manuscript. The contents are organized as follows:

- **Section 1:** Detailed construction process and comprehensive degradation analysis of our WildDark-Face dataset.
- **Section 2:** Additional ablation studies on hyperparameter robustness, as well as inference efficiency and flexibility.
- **Section 3:** Calculation details of our user study.
- **Section 4:** Additional visual ablation studies and extensive qualitative comparisons against existing methods.

1 Construction and Degradation Analysis of the WildDark-Face Dataset

To complement the descriptions in the main manuscript, we provide a detailed elaboration on the construction process, as well as a comprehensive statistical and visual analysis of the proposed WildDark-Face benchmark. This dataset comprises 700 unconstrained real-world facial images that suffer from severe, compound degradations.

*Equal contribution.

†Corresponding author: Guangwei Gao (gwgao@njjust.edu.cn)

1.1 Dataset Construction

Our WildDark-Face dataset is meticulously derived from the widely recognized DarkFace benchmark. Notably, while the extremely low-light conditions of these unconstrained scenes severely obscure human vision, modern robust face detectors are still capable of successfully localizing the facial regions, yielding accurate bounding box coordinates $(x_{min}, y_{min}, x_{max}, y_{max})$. As illustrated in Fig. S1, we leverage these provided annotations to precisely crop individual faces from the dark raw environments. To ensure the evaluation reliability of the benchmark for restoration tasks, we execute a rigorous curation process, mathematically filtering out cropped instances with extremely low spatial resolutions or invalid boundary coordinates.

However, this extraction process exposes a critical real-world bottleneck. Although these faces can be successfully detected, the extracted facial crops are submerged in extreme darkness. They exhibit severely blurred contours, profound sensor noise, and complex compound degradations. Consequently, while these faces are spatially localized, they are visually imperceptible and lack the structural integrity required for downstream identity-aware applications, leading to highly unreliable results in face matching and recognition tasks. This stark contrast between *high detectability* and *poor recognizability* directly motivates our work. It highlights that merely detecting faces in the dark is far from sufficient; rather, our PASDiff bridges this gap by employing a multi-objective guidance strategy to recover the photometric information and high-frequency facial semantics, thereby effectively enhancing the recognizability of these degraded faces for subsequent tasks.

1.2 Degradation Analysis

Quantitative Degradation Distributions. We quantitatively analyze the degradation severity of WildDark-Face across three critical dimensions, as illustrated in Fig. S2:

- *Illumination Intensity*: Fig. S2 (**left**) demonstrates that the average pixel intensities are heavily compressed into the extremely dark region (predominantly between 0 and 20). This confirms the extreme photon starvation present in the dataset.
- *Color Degradation*: We robustly measure color degradation using the average chroma $(\sqrt{a^2 + b^2})$ in the perceptually uniform CIELAB color space. As shown in Fig. S2 (**middle**), the distribution is densely concentrated at low values, quantitatively verifying the severe color desaturation.
- *Structural Blurriness*: Fig. S2 (**right**) reports the structural blurriness, measured by the variance of the Laplacian (after standard Gaussian denoising to prevent high-frequency noise interference). The remarkably low variance scores demonstrate the severe loss of high-frequency structural edges caused by unpredictable motion and out-of-focus blur.



Fig.S1: Visual example of the annotated bounding boxes in the original full-scene image. These bounding boxes are localized by an existing face detector. Based on the provided coordinate information, we precisely crop these regions to systematically extract individual facial images.

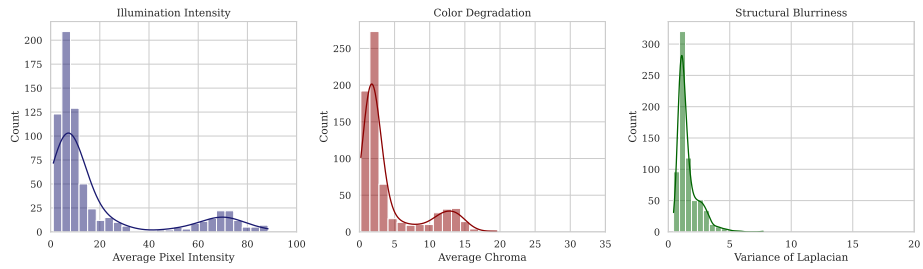


Fig.S2: Statistical degradation distributions of the proposed WildDark-Face benchmark, demonstrating extreme photon starvation, color fading, and severe structural blurriness.

Visual Examples of Compound Degradations. While statistical distributions provide a macroscopic view, we further present a qualitative gallery in Fig. S3 to intuitively showcase the dataset. We arrange the representative visual samples into three rows:

- *Row 1 (Original Inputs):* Demonstrates the extreme darkness of the raw captures, where facial structures and identities are nearly imperceptible.
- *Row 2 (Concealed Degradations):* Displays naively brightened versions of the raw inputs. This directly exposes the severe compound degradations hidden in the dark, such as overwhelming sensor noise and intractable motion blur.
- *Row 3 (Unconstrained Poses):* Presents naively brightened instances with diverse head poses, illustrating the unconstrained geometric conditions of the real-world captured faces.



Fig. S3: Representative samples from the WildDark-Face dataset. The first row shows the original extremely dark inputs. The second and third rows display naively brightened versions, exposing severe concealed degradations (noise and blur) and diverse unconstrained head poses, respectively.

2 Additional Ablation Studies and Inference Efficiency

2.1 Additional Ablation Studies

Robustness to Hyperparameter Variations. While the main manuscript extensively validates the indispensable role of each proposed loss component, we further investigate the hyperparameter robustness of our PASDiff framework in this section. Specifically, we evaluate the visual stability of our framework across two critical constraints: the exposure constraint weight (λ_{exp}) and the structural

injection weight (λ_{stru}). We perturb λ_{exp} by ± 200 around its default setting of 1200 (i.e., testing at 1000, 1200, and 1400). Similarly, we adjust λ_{stru} by ± 2000 from its default configuration of 10000 (i.e., testing at 8000, 10000, and 12000).

Qualitative results are presented in Fig. S4. As observed, altering λ_{exp} safely maintains the overall illumination level without causing under-exposure or over-exposure. Likewise, varying λ_{stru} does not lead to structural collapse, residual blur, or over-sharpening artifacts; the identity characteristics and high-frequency details remain consistently stable. The visual consistency across these weight variations demonstrates that the robustness of PASDiff is primarily attributed to the effective decoupling of photometric correction and structural refinement, rather than being overly dependent on specific hyperparameter settings.



Fig. S4: Visual evaluation of hyperparameter robustness. The top row illustrates the consistency across different exposure weights (λ_{exp}). The bottom row demonstrates the stability of our model under varying structural weights (λ_{stru}). PASDiff maintains clear structural details and natural illumination without requiring meticulous parameter tuning.

2.2 Inference Efficiency and Flexibility

Furthermore, we evaluate the inference efficiency of our proposed framework compared to the most competitive diffusion-based cascaded baselines. The average inference time is measured on a single NVIDIA RTX 3090 GPU with an input resolution of 256×256 .

Cascaded approaches inherently compound the computational burden, as they require executing an enhancement model (e.g., LightenDiffusion) followed

by a separate diffusion restoration model. For instance, while coupling LightenDiffusion with TSD-SR achieves fast inference, it struggles with severe perceptual distortion and structural misalignment. Conversely, employing DiffBIR as the restorer yields better perceptual quality but significantly increases the computational cost. In contrast, PASDiff unifies both tasks within a single reverse diffusion trajectory. A key advantage of our framework is its inherent flexibility: the number of gradient optimization steps (N) per denoising interval can serve as a controller for the speed-quality trade-off. While larger N yields the ultimate restoration fidelity reported in the main manuscript, users can dial down this parameter to prioritize inference speed.

To demonstrate this, we evaluate an accelerated configuration of our framework with a step count ($N = 2$). As detailed in Tab. S1, this lightweight setting achieves an average inference time of 4.92s, which is more efficient than the top-performing cascaded pipeline (LightenDiffusion→DiffBIR, 5.41s). Notably, despite the aggressive reduction in iterative optimization, our $N = 2$ configuration consistently outperforms both baselines in crucial perceptual metrics (LPIPS and Deg.) and facial fidelity (LMD). This validates that PASDiff can flexibly adapt to efficiency requirements while robustly preserving natural illuminations and facial semantics.

Table S1: Quantitative comparison of inference efficiency and restoration performance. To expedite the evaluation process, we randomly selected 100 inputs to weigh the efficiency of our model. Our method achieves a favorable balance between restoration quality and inference speed.

Method	Inference Time (s) ↓	LPIPS ↓	Deg. ↓	LMD ↓
LightenDiffusion → TSD-SR	1.24	0.4126	9.0737	2.6448
LightenDiffusion → DiffBIR	5.41	0.3063	7.2990	2.1906
Ours	4.92	0.2957	7.2406	1.6155

3 Details of the User Study

To complement the subjective preference results presented in the main manuscript, we provide the detailed protocol, experimental setup, and calculation methodology of our user study in this section.

Participants and Dataset. We invited 10 expert evaluators, comprising researchers and practitioners familiar with image restoration tasks, to participate in the subjective assessment. The evaluation set consisted of 20 challenging images (10 synthetic and 10 real-world captures) randomly sampled from our test datasets to avoid selection bias.

Evaluation Protocol. We adopted a rigorous blind forced-choice ranking mechanism. For each of the 20 test cases, evaluators were presented with the degraded

input alongside the restored outputs from PASDiff and the top three competing baselines. To prevent visual bias, all comparative images were anonymized, and their spatial display order was completely randomized for every trial. The evaluators were asked to carefully rank the four methods from 1st (best) to 4th (worst) across three independent perceptual dimensions: *natural lighting*, *chromatic realism*, and *structural clarity*.

Score Aggregation and Statistics. To quantify the subjective preferences, we assigned a linear scoring weight to the ranking results: 4 points for 1st place, 3 points for 2nd, 2 points for 3rd, and 1 point for 4th place. With 10 evaluators assessing 20 images, we collected exactly 200 independent evaluation trials for each of the three dimensions.

The final Mean Ranking Score (MRS) for a specific method m on a given dimension d is calculated as the average of the accumulated points over all 200 trials:

$$\text{MRS}_{m,d} = \frac{1}{200} \sum_{i=1}^{10} \sum_{j=1}^{20} \text{Score}(E_i, I_j, m, d), \quad (1)$$

where $\text{Score}(E_i, I_j, m, d) \in \{1, 2, 3, 4\}$ denotes the assigned score from evaluator E_i for image I_j using method m on dimension d . This extensive accumulation of 200 independent trials per dimension ensures that the subjective preferences reported in our main paper are statistically robust and reliably reflect human visual perception.

4 Extensive Qualitative Comparisons

In this section, we provide additional visual results to complement the main manuscript. Specifically, we present more qualitative examples for our ablation studies to validate each proposed component. Furthermore, we provide extensive visual comparisons against the evaluated methods reported in the main paper across synthetic and real-world datasets.

4.1 Additional Visual Ablation Studies

We first present additional visual ablation results corresponding to the three key designs of our framework.

Physical Constraints. Fig. S5 provides more examples demonstrating the impact of the exposure loss ($\mathcal{L}_{exp}$) and the reflectance loss ($\mathcal{L}_{ref}$). Consistent with the main text, removing $\mathcal{L}_{exp}$ leads to reduced visibility, while removing $\mathcal{L}_{ref}$ causes profound color distortion.

Structural Injection. Fig. S6 shows additional results regarding the structural guidance ($\mathcal{L}_{stru}$). Without it, the model struggles with residual blur and fails to preserve high-frequency facial details.

Style-Agnostic Design. Fig. S7 illustrates more cases validating our SASI module. Replacing SASI with a naive MSE loss causes distinct global color shifts, whereas our design effectively distills structures while maintaining the correct physical atmosphere.



Fig. S5: Visual ablation on physical constraints.



Fig. S6: Visual ablation on structural injection.



Fig. S7: Visual ablation on style-agnostic design.

4.2 Visual Comparisons with Existing Methods

To demonstrate the capability of PASDiff, we provide visual comparisons against existing methods. We first present results on synthetic and real-world datasets in Fig. S8 and Fig. S9, comparing our approach against the baselines whose visual results are omitted from the main text. Furthermore, we present comprehensive comparisons against all 10 evaluated baselines (including cascaded models and joint restoration models) in Fig. S10 and Fig. S11. In these evaluations, PASDiff achieves clear structures and natural colors.

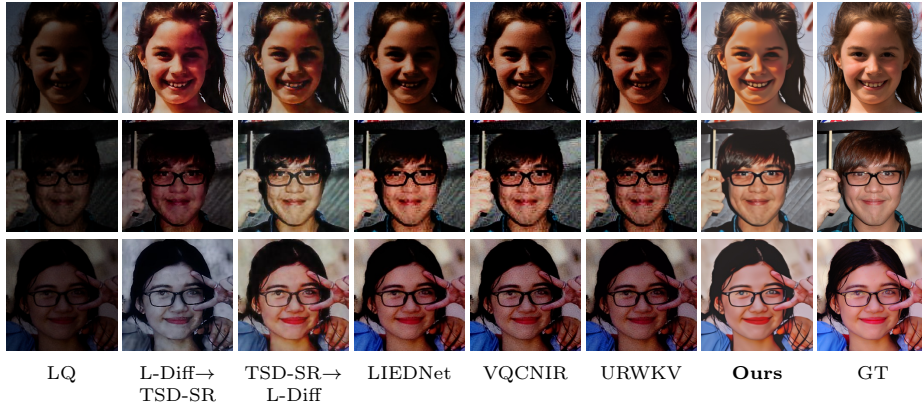


Fig. S8: Supplementary visual comparisons on the synthetic FFHQ dataset.

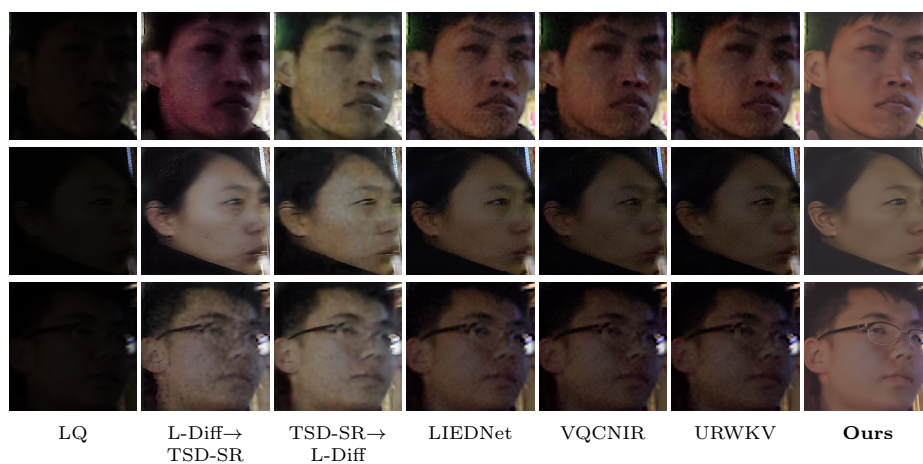


Fig. S9: Supplementary visual comparisons on the real-world WildDark-Face dataset.

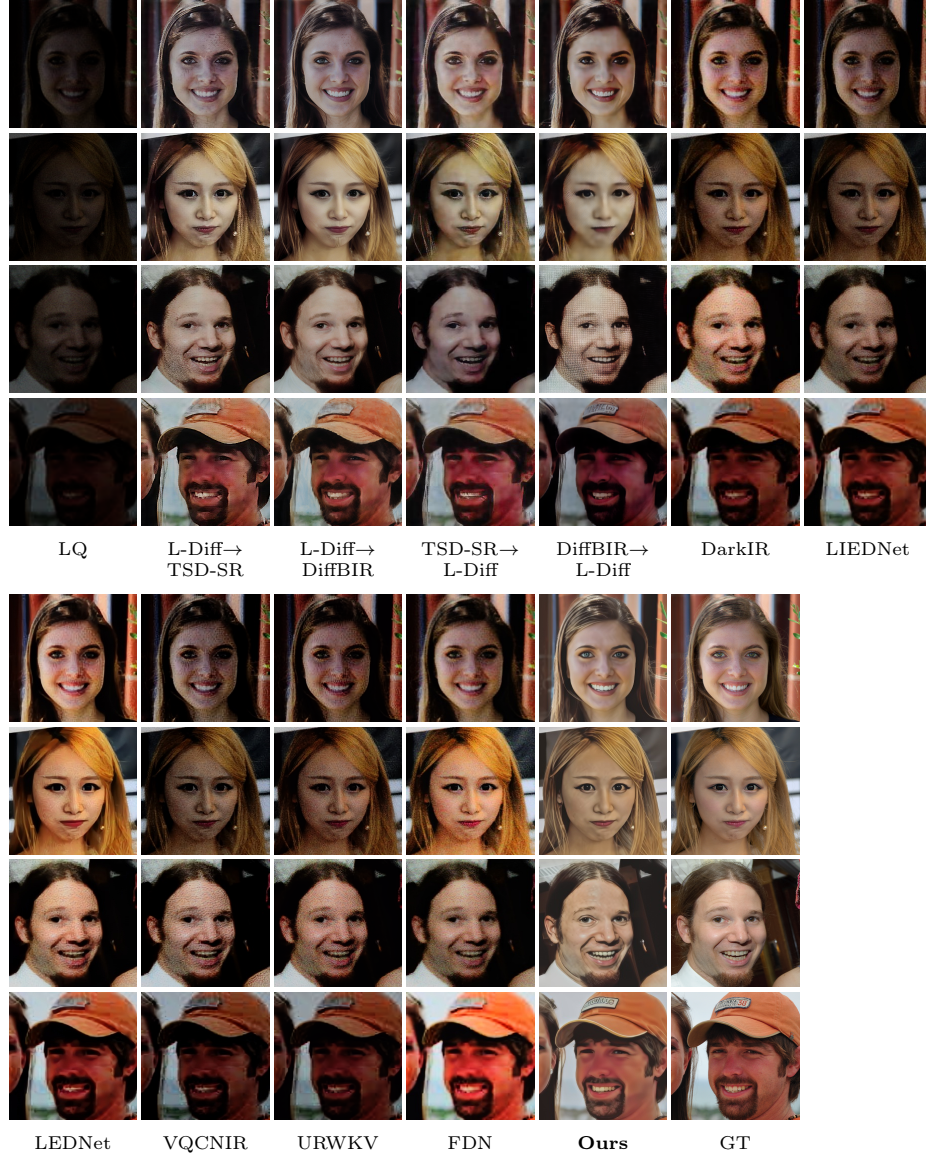


Fig. S10: Comprehensive visual comparisons on the synthetic FFHQ dataset.



Fig. S11: Comprehensive visual comparisons on the real-world WildDark-Face dataset.