DPFR-LLIE: DUAL-PRIOR FREQUENCY ROUTING FOR UAV LOW-LIGHT IMAGE ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

UAV images captured from high altitudes and long viewing distances often suffer from uneven illumination, low signal-to-noise ratios, color distortion, and platform motion. These coupled degradations suppress object boundaries and textures, leaving the degraded observation with insufficient evidence to reliably recover the structural details of small and distant targets. Existing low-light enhancement methods infer restoration cues mainly from this incomplete observation, therefore weak structural information can be mistaken for noise or blur and smoothed away. To obtain structural guidance that generalizes across low-light degradations, we analyze frozen DINOv3 feature representations learned from large-scale visual pretraining. We find that they retain cross-exposure structural correspondence under severe degradation, with low-, mid-, and high-frequency components showing their strongest correspondence at different feature depths. Based on this feature-depth–frequency dependency, we propose Dual-Prior Frequency Routing for Low-Light Image Enhancement (DPFR-LLIE). We first design a frequency-specific feature-depth aggregation module to learn band-specific mixtures of frozen DINOv3 features. Then we design a bounded input-conditioned routing module to allocate and inject the resulting structural guidance with controlled magnitude. Finally, we propose a photometric conditioning module to modulate encoder features with image-derived luminance and chrominance cues. Across the UAV scenes, DPFR-LLIE achieves the best quantitative scores and the strongest qualitative results among the evaluated methods. Specifically, compared with the current SOTA method StarIR, DPFR-LLIE improves PSNR/SSIM by 0.71 dB/0.014 on Synthetic VisDrone Low-Light and by 1.42 dB/0.022 on LOL-Blur. With the same frozen detector, the enhanced UAV images further improve AP and AP_S by 0.27 and 0.31 points. The code will be released upon acceptance.

1 INTRODUCTION

Low-light image enhancement for unmanned aerial vehicles (UAVs) aims to recover visibility from images degraded by uneven illumination, color distortion, noise, and blur (Cai et al., 2023; Xu et al., 2022). In aerial scenes, small and distant objects may occupy only a few pixels, making their structures highly sensitive to degradation (Zhang et al., 2019). Enhancement therefore needs to correct photometric distortions while preserving weak structural details.

Many existing methods derive restoration cues from the degraded image itself, for example by modeling illumination or estimating signal quality (Cai et al., 2023; Xu et al., 2022). Under severe UAV degradation, the observed pixels may provide too little reliable evidence to distinguish faint object boundaries from noise or blur. Such cues can improve visibility while faint structures remain indistinct. This motivates structural guidance learned beyond the degraded input.

Pretrained visual encoders provide structural representations learned beyond the restoration domain (Lin et al., 2023; Luo et al., 2024), but their pretraining objectives do not explicitly target correspondence between severely degraded images and well-exposed references. Noise and blur can obscure weak boundaries, so directly transferred features do not guarantee reliable structural guidance for UAV LLIE. Prior work has addressed this structural-guidance problem through different strategies. For example, SAM-based methods derive structural cues from segmentation region (Li et al., 2024).Although SAM-based enhancement has achieved strong performance by deriving structural guidance from segmentation regions (Li et al., 2024), its reliability remains tied to the quality of those regions. In low-light UAV imagery, noise and blur can obscure object bound-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

aries, making segmentation-derived guidance unstable and limiting the reliability of direct feature transfer. DINOv3 offers a complementary advantage through large-scale self-supervised learning and Gram anchoring (Siméoni et al., 2025). These mechanisms promote spatially coherent and semantically consistent feature maps, enabling the model to better capture inter-patch relationships and cross-exposure correspondences. DINOv3 may therefore provide more robust structural guidance in low-light scenes, motivating us to examine its cross-exposure correspondence rather than assume low-light robustness from feature transfer alone.

We analyze frozen DINOv3 token maps through low-, mid-, and high-frequency decomposition. Their spectral composition changes with feature depth, and their cross-exposure correspondence varies across frequency bands and degradation levels. Under extreme degradation, low-frequency correspondence peaks at a different feature depth from mid- and high-frequency correspondence. This feature-depth–frequency heterogeneity motivates separate feature-depth mixtures for different bands rather than a single feature depth or one shared mixture. Correspondence characterizes feature consistency. We evaluate the resulting construction and allocation strategies separately through controlled restoration comparisons.

Based on these observations, we propose **DPFR-LLIE**, a UAV low-light enhancement framework with two structural-guidance modules and a photometric conditioning module. We first design a frequency-specific feature-depth aggregation module that learns an image-independent feature-depth mixture within each band. We then design a bounded input-conditioned routing module that allocates the guidance branches around a uniform anchor. Equal-RMS control matches the aggregate residual magnitude to the uniform reference, limiting residual-scale differences. Finally, we design a photometric conditioning module that modulates three encoder stages with image-derived luminance and chrominance cues. The routed guidance enters the coarsest decoder stage.

Comparisons across one or multiple feature depths assess guidance construction, while Uniform, Static, and Dynamic injection assess allocation. Dual-prior ablations test integration with photometric conditioning. Together, these controls separate structural-prior design from complete-system performance.

Our contributions are as follows:

- We characterize feature-depth–frequency heterogeneity in frozen DINOv3 under UAV low-light degradation. The observed frequency-dependent correspondence peaks motivate a frequency-specific formulation across feature depths.
- We introduce a two-stage structural-guidance strategy comprising a frequency-specific feature-depth aggregation module followed by a bounded input-conditioned routing module. Distinct feature-depth mixtures construct the guidance, while bounded mixing and Equal-RMS control regulate its allocation and aggregate residual magnitude.
- We integrate our photometric conditioning module with the routed structural guidance and evaluate the resulting framework on Synthetic VisDrone, LOL-Blur, and transfer to DarkRS, including comparisons of guidance construction, injection policies, and prior complementarity.

2 Related Work

Low-Light Image Enhancement. Low-light image enhancement (LLIE) aims to restore visibility, color fidelity, and structural details from images captured under limited illumination. Existing methods introduce various image priors to address the inherent ambiguity of this restoration problem. SNR-Aware LLIE exploits signal-to-noise information to distinguish reliable and degraded regions (Xu et al., 2022), while Retinexformer models illumination components to guide enhancement (Cai et al., 2023). Other methods incorporate structural priors (Xu et al., 2023) or specialized color representations (Yan et al., 2025). Li et al. derive image-level edges from SAM segmentation and region-aware feature priors to guide structural recovery and illumination modeling across regions (Li et al., 2024). Multinex integrates multiple analytic illumination and color representations into a Retinex residual formulation (Brateanu et al., 2026). These studies show how structural and photometric priors guide different aspects of LLIE. Different from the above methods, DPFR-LLIE integrates our photometric conditioning module with structural guidance constructed from frozen DINOv3 features. Our focus is how to organize that guidance across feature depths and frequency bands and allocate it for different inputs.

Foundation-Model Priors and Conditional Restoration. Pretrained visual models have recently been introduced into image restoration as additional sources of structural and semantic guidance. DINO-IR combines shallow and deep DINOv2 features to capture both local details and high-level

information (Lin et al., 2023), while DA-CLIP adapts pretrained CLIP representations to characterize degradation conditions for conditional restoration (Luo et al., 2024). AirNet (Li et al., 2022) and PromptIR (Potlapalli et al., 2023) condition shared restoration networks using learned degradation representations or prompts. For illumination restoration, DINOLight adaptively combines features from multiple DINOv2 layers through spatial weight maps and integrates them using spatial- and frequency-domain attention (Oh et al., 2026a). CANDLE selects and fuses multi-layer DINOv3 features across encoder stages and introduces color-frequency refinement for ambient lighting normalization (Jian et al., 2026). These methods highlight the value of multi-layer foundation-model features for restoration. Different from the above methods, DPFR-LLIE examines how degraded-to-reference feature correspondence varies with feature depth, frequency band, and degradation severity. This analysis motivates a separate feature-depth mixture for each band, followed by bounded input-conditioned allocation of the resulting guidance branches.

Frequency-Aware Image Restoration. Frequency-domain representations provide a useful perspective for separating image components with different restoration behaviors. FFTformer performs attention and feed-forward operations in the frequency domain for image deblurring (Kong et al., 2023), while FourLLIE combines Fourier amplitude estimation with spatial-domain processing for low-light enhancement (Wang et al., 2023). AdaIR adaptively decomposes frequency information according to degradation conditions and modulates low- and high-frequency components for all-in-one restoration (Cui et al., 2025). OmniLight combines DINO-guided restoration with a wavelet-domain mixture-of-experts framework, where degradation-aware routing processes low- and high-frequency sub-bands (Oh et al., 2026b). In comparison, DPFR-LLIE applies smooth Fourier masks to projected token maps extracted from multiple frozen DINOv3 feature depths, generating separate low-, mid-, and high-frequency components. Features are first aggregated across feature depths within each frequency band, followed by an input-conditioned router that determines the combination of the three guidance branches. Uniform and learned static mixtures provide control settings for comparing restoration outcomes under different injection policies.

3 METHOD

DPFR-LLIE comprises a three-level restoration encoder–decoder with photometric conditioning and routed structural guidance from frozen DINOv3. The photometric conditioning module modulates encoder features with luminance and chrominance cues. Frequency-specific feature-depth aggregation and bounded input-conditioned routing construct and inject structural guidance before the coarsest decoder stage. We first examine the feature-depth–frequency behavior, then detail the architecture and training objective.

3.1 EMPIRICAL FEATURE-DEPTH–FREQUENCY ANALYSIS

We analyze all 12 blocks of the frozen DINOv3 ViT-S/16 with 224×224 inputs and $384 \times 14 \times 14$ token maps. Smooth radial Fourier masks define low-, mid-, and high-frequency bands. Figure 1 summarizes spectral composition across 2,059 Synthetic VisDrone validation images and LQ–GT cosine correspondence under light and extreme degradation. Appendix A provides the complete measurement protocol.

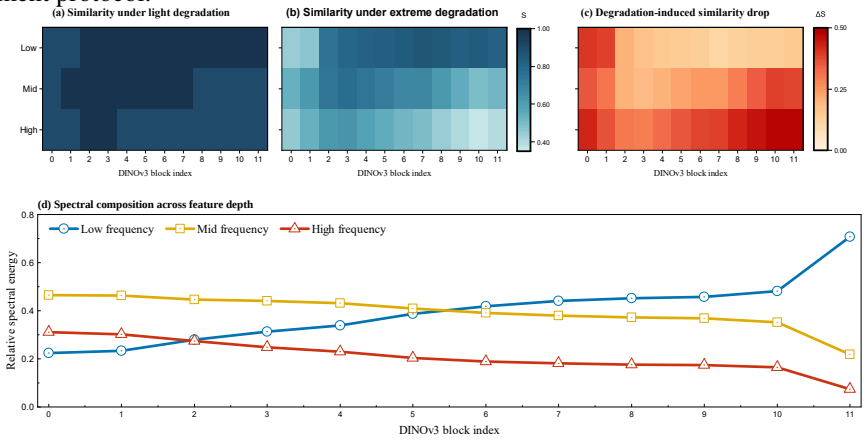

Figure 1: Feature-depth–frequency analysis of frozen DINOv3: (a,b) LQ–GT correspondence under light and extreme degradation, (c) the group difference, (d) spectral composition. Appendix A gives the protocol.

Spectral composition varies across feature depths. The low-frequency spectral fraction increases from 0.224 at block 0 to 0.707 at block 11, while the mid- and high-frequency fractions decrease from $0.466/0.310$ to $0.219/0.074$. Therefore, different DINOv3 feature depths provide distinct spectral compositions under the same frequency decomposition.

Feature correspondence changes with frequency and degradation severity. The mean low/mid/high LQ–GT cosine correspondence is $0.955/0.923/0.888$ in the light degradation setting and $0.766/0.639/0.500$ in the extreme setting. The larger changes in mid- and high-frequency correspondence suggest that different bands should be considered separately. Since the two settings contain different samples, these comparisons describe group-level differences rather than paired degradation changes.

Different frequency bands favor different feature depths. Under extreme degradation, correspondence peaks at block 7 for low frequencies and block 3 for mid and high frequencies. These band-dependent preferences motivate separate feature-depth mixtures. We select regularly spaced blocks $\{0, 2, 4, 6, 8, 10\}$ as a common candidate set and learn their contributions for each band. Appendix A explains the distinction between correspondence peaks and this feature-depth sampling choice. Section 4.3 compares the resulting guidance and injection configurations.

3.2 Dual-Prior Guidance Modules

Given a low-light image $x \in \mathbb{R}^{3\times H\times W}$, our framework comprises a three-level encoder–decoder restoration network that predicts $\hat{y}$. A 3×3 convolution embeds x before encoder stages E_1, E_2, E_3, which contain 1, 2, and 4 restoration blocks with 24, 48, and 96 channels, respectively. The same x also enters our photometric conditioning module in Figure 2(c) and the frozen DINOv3 branch in Figure 2(b). Photometric affine parameters condition the three encoder levels, while the structural-guidance residual enters after E_3 and immediately before decoder stage D_3. The decoder incorporates two encoder skip fusions, followed by refinement and a 3×3 output convolution to predict $\hat{y} = x + \Delta x$. At inference, both guidance pathways depend only on x. The paired reference y provides training supervision only. Figure 2 shows the complete pipeline and module details.

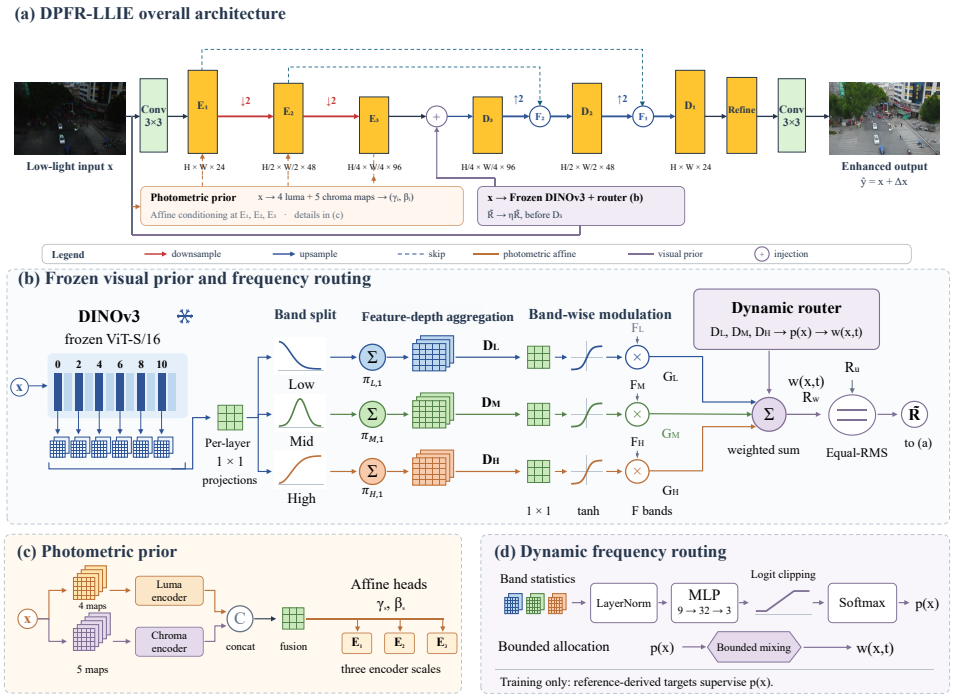

Figure 2: DPFR-LLIE framework. (a) Restoration path with photometric encoder conditioning and structural decoder guidance. (b) Frozen DINOv3 features undergo projection, frequency decomposition, feature-depth aggregation, routing, and Equal-RMS control. (c) Photometric module. (d) Bounded router with training-only reference supervision. Feature planes and frequency curves are schematic.

Photometric conditioning module. Inspired by the multiple analytic luminance and chrominance representations in Multinex (Brateanu et al., 2026), we design a photometric conditioning module for the encoder. It computes four luminance maps and five chrominance maps from x and encodes them into a 32-channel feature representation $V(x)$. Figure 2(c) shows the two encoding branches and three affine heads. Let E_s denote the feature entering the block stack at encoder level s after the initial embedding or the preceding downsampling operation. An affine head predicts $[\gamma_s, \beta_s] = \mathcal{A}_s(\mathcal{U}_s(V(x)))$, where $\mathcal{U}_s$ resizes the photometric feature to that level. The conditioned feature is

$$\widetilde{E}_s = E_s \odot [1 + 0.1 \cdot \tanh(\gamma_s)] + 0.1 \cdot \tanh(\beta_s). \tag{1}$$

The three levels contain 24, 48, and 96 channels, respectively. Zero-initialized output convolutions ensure that the modulation starts from an identity mapping. Appendix B provides details of the analytic maps and encoder design.

Frequency-specific feature-depth aggregation module. Our module extracts normalized patch-token maps $Z_l(x)$ from zero-based DINOv3 blocks $\mathcal{S} = \{0, 2, 4, 6, 8, 10\}$. Each trainable bias-free 1×1 projection P_l reduces 384 channels to 24 channels. For $b \in \{L, M, H\}$, the frequency decomposition is defined as

$$\mathcal{B}_b(A) = \mathrm{irFFT2}(M_b \odot \mathrm{rFFT2}(A)) \,. \tag{2}$$

The orthonormal transforms apply smooth radial masks with transition centers at $1/3$ and $2/3$ and width 0.04, where $\sum_b M_b = 1$. Each frequency band learns an independent feature-depth mixture:

$$\pi_{b,l} = \frac{e^{\theta_{b,l}}}{\sum_{k \in \mathcal{S}} e^{\theta_{b,k}}}, \qquad D_b(x) = \mathcal{U}\left(\sum_{l \in \mathcal{S}} \pi_{b,l} \mathcal{B}_b(P_l Z_l(x)) \right). \tag{3}$$

Projection, decomposition, and aggregation are performed at 14×14 resolution before bilinear resizing $\mathcal{U}$ to the coarsest decoder resolution. Zero-initialized θ provides uniform feature-depth weights at initialization. These weights are shared across images, while input-specific adaptation is introduced by the following frequency router.

Let $F \in \mathbb{R}^{96 \times H/4 \times W/4}$ be the output of E_3 before the structural-guidance injection shown in Figure 2(a). Band-specific 1×1 adapters C_b generate

$$G_b = \tanh(C_b(D_b(x))) \odot \mathcal{B}_b(F). \tag{4}$$

The two streams operate on corresponding relative frequency regions in their native resolutions. The spatial multiplication makes G_b a frequency-conditioned contribution rather than a strictly band-limited output.

Input-conditioned routing module. Figure 2(d) illustrates our predictor and bounded mixing process. The router summarizes each D_b through its RMS magnitude, mean absolute magnitude, and population standard deviation. The logarithms of these nine statistics are concatenated and normalized by LayerNorm to obtain $d(x)$. A $9 \to 32 \to 3$ SiLU MLP predicts

$$p(x) = \mathrm{softmax}(\mathrm{clip}(f_\phi(d(x)), -3, 3)/1.5) \,. \tag{5}$$

The injection weights are initialized around the uniform distribution $u = (1/3, 1/3, 1/3)$:

$$\begin{aligned} c(x) &= 1 - \mathcal{H}(p(x))/\log 3, \qquad \lambda_t = 0.2 \min(t/30000, 1), \\ w(x,t) &= u + \lambda_t c(x)[p(x) - u], \quad \mathcal{H}(p) = -\sum_b p_b \log p_b. \end{aligned} \tag{6}$$

Here, t denotes the number of completed optimizer updates. The injection starts from uniform weights and gradually increases the correction strength over 30k updates. The entropy concentration term $c(x)$ reduces the adjustment magnitude for near-uniform predictions. Since $0 \le \lambda_t c(x) \le 0.2$, the injection weights satisfy $4/15 \le w_b \le 7/15$, ensuring that all branches remain active. Therefore, p determines the correction direction, while w defines the mixture applied during restoration.

Magnitude-controlled injection. Let $R_w = \sum_b w_b G_b$, $R_u = \frac{1}{3} \sum_b G_b$, and $\rho(R) = \sqrt{\langle R^2 \rangle + 10^{-12}}$. The magnitude-controlled residual and injection are defined as

$$\widetilde{R} = \mathrm{sg}\left(\frac{\rho(R_u)}{\max(\rho(R_w), 10^{-12})} \right) R_w, \qquad F' = F + 0.1\sigma(\xi)\widetilde{R}, \tag{7}$$

where sg stops gradients and ξ is initialized to zero. When $\rho(R_w) \le \max(10^{-4}\rho(R_u), 10^{-12})$, $\widetilde{R}$ is set to R_u to avoid unstable rescaling. When $\lambda_t = 0$, $\widetilde{R} = R_u$ directly. Equal-RMS matching controls the overall residual magnitude without normalizing individual branches. The modulated feature F' enters D_3. After D_3, the first upsampling is fused with the E_2 skip at F_2 before D_2. The second is fused with the E_1 skip at F_1 before D_1. Four refinement blocks and the final 3×3 convolution produce Δx, yielding $\hat{y} = x + \Delta x$.

Reference-derived routing supervision. For native-resolution training pairs, the relative RGB frequency discrepancy is

$$e_b = \frac{\langle |\mathcal{B}_b(y) - \mathcal{B}_b(x)| \rangle}{\langle |\mathcal{B}_b(y)| \rangle + 10^{-3}}, \qquad q = \mathrm{sg}[\mathrm{softmax}((e - \bar{e}\mathbf{1})/0.25)], \tag{8}$$

where $\bar{e} = \frac{1}{3}\sum_b e_b$. We minimize $\mathcal{L}_{\mathrm{route}} = -\mathbb{E}_{x,y}\sum_b q_b \log p_b(x)$. The target represents relative restoration demand rather than an optimal routing label or a DINO feature reliability score. Reference images contribute only during training and are unnecessary for predicting $p(x)$.

3.3 Training and Inference

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{pix}} + 0.1\mathcal{L}_{\mathrm{fft}} + 0.1\mathcal{L}_{\mathrm{align}} + 0.05\mathcal{L}_{\mathrm{route}}. \tag{9}$$

$\mathcal{L}_{\mathrm{pix}}$ denotes the image L1 loss. $\mathcal{L}_{\mathrm{fft}}$ computes the average L1 difference between concatenated real and imaginary components of the unnormalized full 2D FFT. For native-resolution band aggregates $\bar{D}_b$, alignment is defined as

$$\mathcal{L}_{\mathrm{align}} = \mathbb{E}_{x,y}\frac{1}{3}\sum_b \left[1 - v(\bar{D}_b(x))^\top \mathrm{sg}(v(\bar{D}_b(y)))\right], \tag{10}$$

where v flattens and ℓ_2-normalizes its input. Both paths share projections and feature-depth weights, while the reference aggregate is detached. DINOv3 remains frozen, and the backbone and guidance modules are optimized jointly. Routing supervision updates the predictor and shared guidance features, while reconstruction gradients also propagate through the active router. During inference, both priors and routing are retained with $\lambda = 0.2$, while reference extraction and auxiliary losses are removed. Appendix B provides further details on numerical stabilization and gradient flow.

4 Experiments

4.1 Datasets and Experimental Settings

Synthetic VisDrone Low-Light contains 19,246 training pairs and 2,059 validation pairs generated with inverse DCE degradation, Gaussian noise, and motion blur (Li et al., 2021; Yao et al., 2024). The dataset includes four degradation levels: light, medium, heavy, and extreme. We separately train models on the official LOL-Blur split (Zhou et al., 2022) and directly transfer them to unpaired nighttime UAV images from DarkRS (Yao et al., 2024) without fine-tuning. PSNR/SSIM are reported on paired datasets, while NIQE is reported on DarkRS.

DPFR-LLIE is implemented in PyTorch/BasicSR with 128×128 crops, batch size 32, AdamW optimization, cosine decay over 300k updates, and frozen DINOv3. All ablations share the same split, initialization, encoder checkpoint, training budget, and evaluation protocol.

Table 1: Main results. Synthetic VisDrone and LOL-Blur use separate training protocols, the LOL-Blur models transfer to DarkRS without fine-tuning. Higher PSNR/SSIM and lower NIQE are better.

	Synthetic VisDrone		LOL-Blur		DarkRS
Method	PSNR	SSIM	PSNR	SSIM	NIQE
Zero-DCE	14.58	0.515	14.95	0.612	5.072
SNR-Aware	20.86	0.732	21.87	0.811	9.205
FourLLIE	14.58	0.433	18.53	0.732	4.130
RetinexFormer	24.15	0.809	24.00	0.840	5.440
Multinex	22.06	0.753	22.82	0.823	**3.959**
StarIR	26.10	0.853	26.52	0.905	4.184
DPFR-LLIE	**26.81**	**0.867**	**27.94**	**0.927**	4.129

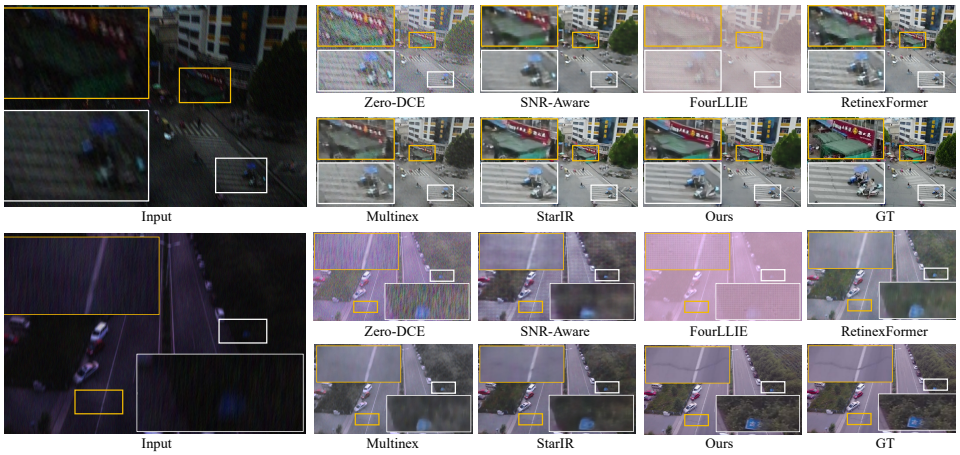

Figure 3: Qualitative comparison on Synthetic VisDrone Low-Light under extreme degradation. Enlarged regions highlight local structures.

4.2 Comparison with Existing Methods

We compare with Zero-DCE (Li et al., 2021), SNR-Aware (Xu et al., 2022), FourLLIE (Wang et al., 2023), RetinexFormer (Cai et al., 2023), Multinex (Brateanu et al., 2026), and StarIR (Cui et al., 2026). Table 1 summarizes the three evaluation settings.

On Synthetic VisDrone, DPFR-LLIE improves on StarIR by 0.71 dB PSNR and 0.014 SSIM. Its PSNR gains at light, medium, heavy, and extreme degradation are 0.93, 0.86, 0.58, and 0.48 dB, both metrics improve at every level (Appendix E). The margin narrows as degradation becomes more severe but remains positive across all four levels. The gain arises because the photometric conditioning module modulates three encoder scales with luminance and chrominance cues to correct exposure and color, while the frequency-specific feature-depth aggregation module aggregates DINOv3 cues across feature depths within each band to help preserve weak boundaries under noise and blur. The dual-prior ablation in Table 2 supports their complementarity. On LOL-Blur, the gains reach 1.42 dB and 0.022 SSIM under a separate training protocol. DarkRS NIQE is 4.129, below StarIR's 4.184 but above Multinex's 3.959, indicating competitive transfer rather than the best no-reference score.

Figure 3 compares two UAV scenes under extreme degradation. DPFR-LLIE recovers clearer vehicle contours, roadside signs, and lane boundaries, while also reducing the color distortions produced by several competing methods. Clearer contours are consistent with band-specific DINOv3 guidance because it enters before decoding, while improved color and exposure are consistent with photometric conditioning because luminance and chrominance cues modulate the encoder. The branch ablation in Table 2 tests these contributions beyond the two visual examples. Appendix E includes LOL-Blur and DarkRS visual comparisons.

4.3 Ablation Studies

We investigate three aspects of DPFR-LLIE: the complementarity of the two priors (Table 2), the input dependence of frequency allocation (Table 3), and the construction of frequency-specific guidance (Table 4). The latter two experiments correspond to the research questions introduced in Section 1.

Table 2: Dual-prior ablation (M: photometric; D: DINO guidance).

Variant	M	D	PSNR↑	SSIM↑
Backbone only	×	×	26.10	0.853
Photometric only	✓	×	26.69	0.865
DINO only	×	✓	26.47	0.866
Full model	✓	✓	**26.81**	**0.867**

Table 3: Injection-policy ablation, the auxiliary predictor remains active.

Variant	Learned	Input-cond.	PSNR↑	SSIM↑
Uniform	×	×	26.73	0.867
Static	✓	×	26.75	**0.868**
Dynamic (full)	✓	✓	**26.81**	0.867

Complementary priors. Table 2 shows that photometric conditioning raises the backbone from 26.10/0.853 to 26.69/0.865 in PSNR/SSIM, its three-level modulation gives the encoder exposure and color cues before reconstruction. The complete DINO branch reaches 26.47/0.866 by adding structural guidance beyond the degraded pixels. The integrated model reaches 26.81/0.867, consis-

tent with the two branches addressing different sources of error. DINO-only includes alignment and routing objectives, so its gain cannot be assigned to the frozen encoder alone. Appendix D gives the controls.

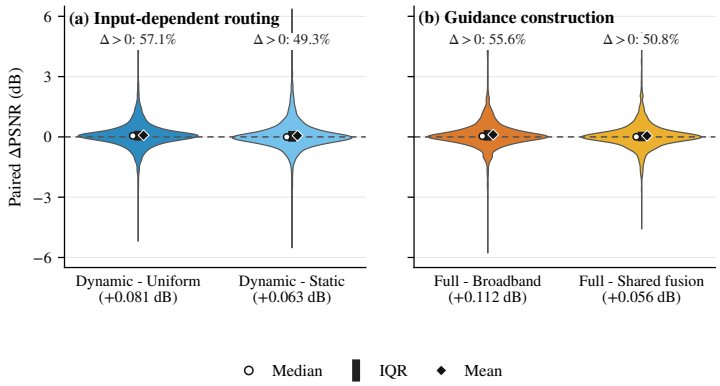

Figure 4: Paired per-image PSNR differences on Synthetic VisDrone: (a) Dynamic versus Uniform/Static, (b) full guidance versus Broadband/Shared. Markers show median, mean, and IQR, labels give the positive ratio and mean difference.

Input-dependent versus static allocation. Table 3 compares equal, globally learned, and input-conditioned injected mixtures while retaining the auxiliary predictor in every variant. Appendix D gives the matched control definitions. Dynamic improves mean PSNR by 0.08/0.06 dB over Uniform/Static, with SSIM equal to Uniform and 0.001 below Static. Because band correspondence varies across degradation groups (Figure 1), one fixed mixture may not suit every input. Dynamic adjusts band weights from each image's statistics, bounded mixing keeps all bands active, and Equal-RMS holds aggregate residual scale comparable. This offers a reason for its modest PSNR gain, while SSIM does not exceed Static. Figure 4 reports the paired per-image differences.

Table 4: Frequency and feature-depth aggregation controls on Synthetic VisDrone. Freq. marks decomposition. Feat. depths counts selected DINO blocks. Mix specifies feature-depth weighting. Single feature depth is block 10.

Variant	Freq.	Feat. depths	Mix	PSNR↑	SSIM↑
Broadband	×	6	Separate	26.70	0.866
Single feature depth	✓	1	–	26.74	0.866
Uniform fusion	✓	6	Uniform	26.72	0.866
Shared fusion	✓	6	Shared	26.75	0.866
Full model	✓	6	Separate	**26.81**	**0.867**

Frequency decomposition and feature-depth aggregation. Table 4 compares broadband, single-feature-depth, uniform-fusion, and shared-fusion guidance construction. Full reaches 26.81 dB versus 26.70 dB for Broadband and 26.75 dB for shared fusion. Because correspondence peaks differ by band (Figure 1), shared fusion must use one feature-depth mixture for all bands, separate mixtures remove that constraint, consistent with the 0.06 dB gain over shared fusion. The 0.11 dB gap to Broadband also changes routing supervision and cannot be assigned to decomposition alone (Appendix D). Single feature depth tests zero-based block 10 only. Taken together, these controls support band-specific construction across feature depths, while Table 3 shows only a modest additional PSNR gain from input-conditioned allocation. Appendix C gives the full protocol.

4.4 Downstream Object Detection

We evaluate every enhancement method with the same frozen Faster R-CNN R50-FPN detector trained for 12 epochs on VisDrone. Only the 2,059 validation inputs change, annotations, preprocessing, checkpoint, and COCO evaluator are fixed. Appendix C.1 gives the complete protocol and discussion.

Table 5: Fixed-detector results on Synthetic VisDrone Low-Light. All values are percentages, higher is better.

Input	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Low input	8.32	15.15	8.11	3.86	13.24	24.59
Zero-DCE	6.26	11.50	6.07	3.09	10.00	16.11
SNR-Aware	8.31	15.28	8.01	3.43	13.13	27.29
FourLLIE	6.98	12.33	6.97	3.47	10.73	16.52
RetinexFormer	11.59	20.91	11.47	5.71	18.18	31.60
Multinex	11.15	19.96	11.07	5.61	17.24	27.86
StarIR	14.85	26.60	14.65	8.10	22.46	35.81
DPFR-LLIE	**15.12**	**26.96**	**14.97**	**8.41**	**22.92**	**36.76**

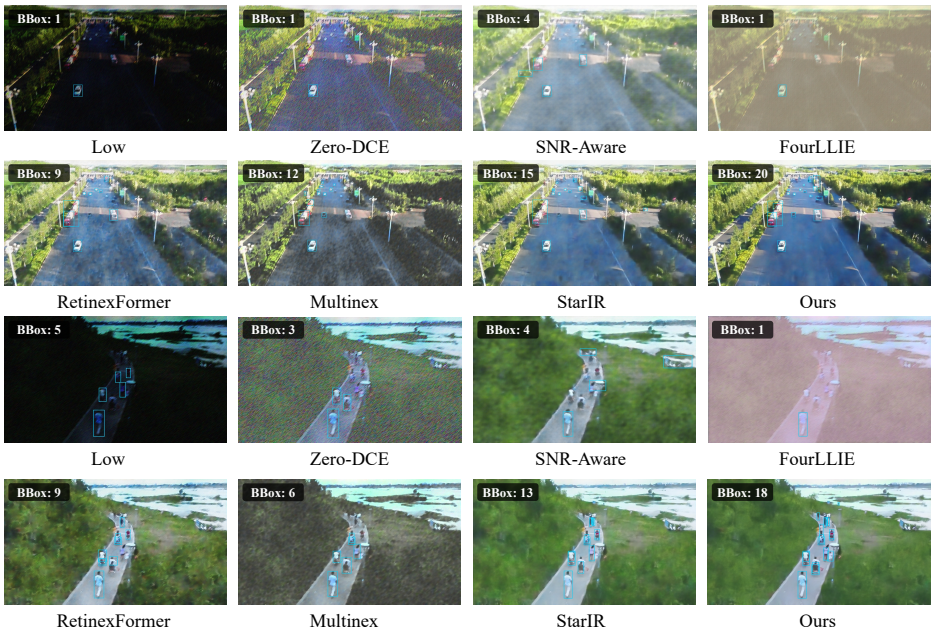

Figure 5: Fixed-detector predictions on two selected scenes.

DPFR-LLIE leads all metrics in Table 5, reaching 15.12 AP and 8.41 AP_S, gains of 0.27 and 0.31 points over StarIR and 6.80 and 4.55 points over the low-light input, respectively. Because the detector and evaluation protocol are fixed, these differences reflect changes in the enhanced images. Clearer faint contours may help the fixed detector distinguish small targets from the background, consistent with the AP_S gain and Figure 3, this test does not isolate either prior. Figure 5 shows cyan-box detector predictions on two selected scenes. Appendix C.1 provides all-method TP/FP/FN visualizations and fixed-operating-point error statistics. Those selected-scene diagnostics are separate from the validation-set COCO AP in Table 5.

5 Conclusion

In this paper, we propose DPFR-LLIE for UAV low-light image enhancement. We analyze feature-depth–frequency heterogeneity in frozen DINOv3, learn a separate feature-depth mixture for each frequency band, and allocate the resulting structural guidance through bounded input-conditioned routing. We integrate this guidance with photometric conditioning to address weak details alongside illumination and color distortions. On Synthetic VisDrone and LOL-Blur, DPFR-LLIE improves paired reconstruction over StarIR. Ablation studies support the complementary roles of photometric conditioning and structural guidance, favor frequency-specific aggregation across feature depths. They also show a modest additional PSNR gain from dynamic routing over static allocation. Enhanced Synthetic VisDrone images also improve AP with a fixed detector. We evaluate no-reference transfer on DarkRS, while paired reconstruction fidelity on real nighttime UAV images remains to be tested.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation to assist with language editing, manuscript restructuring, literature retrieval, and LaTeX formatting checks. All AI-assisted suggestions and modifications were reviewed and verified by the authors. The authors take full responsibility for the final content of the manuscript.

ETHICS STATEMENT

This study uses existing research datasets and does not involve new data collection or experiments with human participants. Low-light enhancement may assist visual analysis under limited illumination. However, benchmark performance should not be interpreted as a guarantee of safety or reliability in surveillance or other safety-critical applications. Practical deployment should consider privacy concerns, dataset bias, potential failure cases, applicable licenses, and relevant legal requirements.

REPRODUCIBILITY STATEMENT

Sections 3–4 describe the proposed model and experimental settings. Appendix A provides the feature-analysis protocol. Appendix B details the complete equations, numerical settings, and gradient flow. Appendix C records dataset splits, optimization details, and fixed-checkpoint evaluation, and Appendix D defines the ablation variants. These materials are intended to support independent implementation and evaluation.

REFERENCES

Alexandru Brateanu, Tingting Mu, Codruta O Ancuti, and Cosmin Ancuti. Multinex: Lightweight low-light image enhancement via multi-prior retinex. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29887–29896, 2026.

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 *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 12470–12479. IEEE, 2023.

Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Khan. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In *International Conference on Learning Representations*, volume 2025, pp. 101306–101327, 2025.

Yuning Cui, Syed Waqas Zamir, Ming-Hsuan Yang, Alois Knoll, Fahad Shahbaz Khan, and Salman Khan. Starir: Convolutional image restoration with spatial-frequency fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Rong-Lin Jian, Ting-Yao Chen, Yu-Fan Lin, Chia-Ming Lee, Fu-En Yang, Yu-Chiang Frank Wang, and Chih-Chung Hsu. Candle: Illumination-invariant semantic priors for color ambient lighting normalization. *arXiv preprint arXiv:2604.02785*, 2026.

Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5886–5895. IEEE, 2023.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17431–17441. IEEE, 2022.

Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(8):4225–4238, 2021.

Guanlin Li, Bin Zhao, and Xuelong Li. Low-light image enhancement with sam-based structure priors and guidance. *IEEE Transactions on Multimedia*, 26:10854–10866, 2024.

Xin Lin, Jingtong Yue, Kelvin CK Chan, Lu Qi, Chao Ren, Jinshan Pan, and Ming-Hsuan Yang. Multi-task image restoration guided by robust dino features. *arXiv preprint arXiv:2312.01677*, 2023.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas Schön. Controlling vision-language models for multi-task image restoration. In *International Conference on Learning Representations*, volume 2024, pp. 16226–16246, 2024.

Youngjin Oh, Junhyeong Kwon, and Nam Ik Cho. Dinolight: Robust ambient light normalization with self-supervised visual prior integration. In *International Conference on Pattern Recognition*, pp. 106–121. Springer, 2026a.

Youngjin Oh, Junyoung Park, Junhyeong Kwon, and Nam Ik Cho. Omnilight: One model to rule all lighting conditions. *arXiv preprint arXiv:2604.15170*, 2026b.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in neural information processing systems*, 36:71275–71293, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Chenxi Wang, Hongjun Wu, and Zhi Jin. Fourllie: Boosting low-light image enhancement by fourier frequency information. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 7459–7469, 2023.

Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 17693–17703. IEEE, 2022.

Xiaogang Xu, Ruixing Wang, and Jiangbo Lu. Low-light image enhancement via structure modeling and guidance. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9893–9903. IEEE, 2023.

Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. Hvi: A new color space for low-light image enhancement. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5678–5687. IEEE, 2025.

Zishu Yao, Guodong Fan, Jinfu Fan, Min Gan, and CL Philip Chen. Spatial–frequency dual-domain feature fusion network for low-light remote sensing image enhancement. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Xindi Zhang, Ebroul Izquierdo, and Krishna Chandramouli. Dense and small object detection in uav vision based on cascade network. In *2019 IEEE/CVF international conference on computer vision workshop (ICCVW)*, pp. 118–126. IEEE, 2019.

Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *European conference on computer vision*, pp. 573–589. Springer, 2022.

A Complete Feature-Depth–Frequency Analysis

Frozen DINOv3 features exhibit behavior dependent on feature depth and frequency under UAV low-light degradation. Figure 1 visualizes the measurements that motivate our frequency-specific guidance design across feature depths.

Analysis protocol. We extract normalized intermediate features from all 12 blocks of a frozen DINOv3 ViT-S/16, indexed from 0 to 11. Each image is bicubically resized to 224×224 and normalized using the encoder's ImageNet statistics. The patch tokens from block l are reshaped into $Z_l \in \mathbb{R}^{384\times14\times14}$.

We define three frequency bands on the native one-sided rFFT grid. Let $f_y = \mathrm{fftfreq}(14)$ and $f_x = \mathrm{rfftfreq}(14)$. The normalized radial coordinate is

$$r(u,v) = \frac{\sqrt{f_y(u)^2 + f_x(v)^2}}{0.5\sqrt{2}}. \tag{11}$$

We construct three soft masks:

$$\begin{aligned} M_L(r) &= 1 - \sigma((r-1/3)/0.04)\,, \\ M_H(r) &= \sigma((r-2/3)/0.04)\,, \\ M_M(r) &= 1 - M_L(r) - M_H(r), \end{aligned} \tag{12}$$

where σ is the sigmoid function. The transition centers are $1/3$ and $2/3$, with a transition width of 0.04. By construction, the masks satisfy $M_L + M_M + M_H = 1$.

Using orthonormally normalized Fourier transforms, the band-specific token maps are

$$\begin{aligned} Z_l^b &= \mathrm{irFFT2}(M_b \odot \mathrm{rFFT2}(Z_l))\,, \\ &\quad b \in \{L, M, H\}. \end{aligned} \tag{13}$$

We analyze spectral composition and cross-exposure correspondence separately. For spectral composition, each channel is standardized over its spatial dimensions:

$$\widetilde{Z}_l = \frac{Z_l - \mu_{hw}(Z_l)}{\sqrt{\mathrm{Var}_{hw}(Z_l) + 10^{-6}}}. \tag{14}$$

The relative spectral energy assigned to band b is

$$\rho_{l,b} = \frac{\sum_{c,u,v} M_b(u,v) \left|\mathrm{rFFT2}(\widetilde{Z}_l)_{cuv}\right|^2}{\sum_{b'\in\{L,M,H\}} \sum_{c,u,v} M_{b'}(u,v) \left|\mathrm{rFFT2}(\widetilde{Z}_l)_{cuv}\right|^2}. \tag{15}$$

Here, relative spectral energy refers to the fraction of mask-weighted spectral power on the one-sided rFFT grid. Figure 1(d) reports the average over 2,059 LQ validation inputs.

For cross-exposure correspondence, we apply Eq. (13) directly to the token maps without spatial standardization and compute

$$S_{l,b} = \cos\big(\mathrm{vec}(Z^b_{l,\mathrm{LQ}}), \mathrm{vec}(Z^b_{l,\mathrm{GT}})\big)\,. \tag{16}$$

We average these similarities independently over 502 light samples and 514 extreme samples, and then compute

$$\Delta S_{l,b} = \bar{S}^{\mathrm{light}}_{l,b} - \bar{S}^{\mathrm{extreme}}_{l,b}. \tag{17}$$

Because the two strata contain different samples, $\Delta S_{l,b}$ measures a group-level difference in correspondence rather than a paired sample-level degradation effect.

Observation 1: Spectral composition varies with feature depth. Figure 1(d) shows that the low-frequency fraction increases from 0.224 at block 0 to 0.707 at block 11, while the mid- and high-frequency fractions decrease from $0.466/0.310$ to $0.219/0.074$. Under this decomposition, early and late blocks therefore exhibit substantially different spectral compositions.

Observation 2: Correspondence differences depend on frequency. When averaged over all 12 blocks, the low/mid/high similarities are $0.955/0.923/0.888$ for the light stratum and $0.766/0.639/0.500$ for the extreme stratum. The larger inter-stratum differences at mid and high frequencies show that correspondence changes differently across the three bands under the evaluated degradation conditions.

Observation 3: Peak-correspondence feature depth varies across bands. Within the extreme stratum, the highest mean LQ–GT correspondence occurs at block 7 for the low-frequency band and at block 3 for the mid- and high-frequency bands, as shown in Fig. 1(b). No single feature depth therefore maximizes this correspondence measure across all three bands.

Design implications. These observations motivate frequency-specific aggregation across feature depths. We first learn one image-independent feature-depth mixture per band, then apply bounded input-conditioned allocation across the resulting guidance branches. Frequency/feature-depth controls and injection-policy comparisons examine these two stages, respectively.

Correspondence peaks and the selected feature depths. The analysis covers all 12 blocks, whereas restoration uses the regularly spaced candidate set $\mathcal{S} = \{0, 2, 4, 6, 8, 10\}$. This fixed sampling of six feature depths provides a common pool of projected features for band-specific aggregation. The correspondence peaks at blocks 3 and 7 motivate different feature-depth mixtures across frequency bands rather than a peak-selection rule. Cross-exposure cosine correspondence characterizes feature consistency, while restoration also depends on the learned projections, modulation, and decoding; the correspondence maximum therefore need not identify the best restoration layer. Within $\mathcal{S}$, each band learns its own combination of feature depths. The ablations assess aggregation within this fixed design through uniform and shared-fusion controls, together with a block-10 single-feature-depth reference.

B Architecture, Numerical Details, and Optimization

B.1 Dual-Prior Restoration Framework

Given a low-light image $x \in \mathbb{R}^{3\times H\times W}$, DPFR-LLIE predicts a normally exposed image $\hat{y}$ with a three-level encoder–decoder restoration network inspired by StarIR (Cui et al., 2026). A 3×3 convolution first embeds x. Encoder stages E_1, E_2, E_3 then contain 1, 2, and 4 restoration blocks with 24, 48, and 96 channels, respectively. The same x enters the photometric branch and the frozen DINOv3 branch. Photometric affine parameters condition all three encoder levels, whereas the DINO-derived visual residual is added after E_3 and immediately before decoder stage D_3. The paired reference y is used only during training. Figure 2(a–d) in the main paper shows the backbone and both prior branches.

Photometric conditioning. Inspired by the analytic luminance and chrominance representations in Multinex (Brateanu et al., 2026), we design a photometric conditioning module that computes four luminance maps and five chrominance maps, as shown in Figure 2(c). The luminance stack includes Rec. 709 luminance, max-RGB intensity, HSL-style lightness, and unnormalized RGB ℓ_2 intensity. The chrominance stack contains zero-offset BT.601 Cb/Cr, normalized red and green chromaticities, and HSV saturation. A denominator floor of 10^{-9} is used when computing chromaticity and saturation.

Let $L(x)$ and $C(x)$ denote the luminance and chrominance stacks. Their encoded features are concatenated and fused:

$$V(x) = \mathcal{M}([\mathcal{E}_L(L(x)), \mathcal{E}_C(C(x))]) \,, \tag{18}$$

where brackets denote channel concatenation. The two encoders and the fusion module use 1×1 and 3×3 convolutions with GroupNorm and GELU, producing a 32-channel prior feature.

At encoder level $s \in \{1, 2, 3\}$, a scale-specific head predicts

$$[\gamma_s, \beta_s] = \mathcal{A}_s(\mathcal{U}_s(V(x))), \tag{19}$$

where $\mathcal{U}_s$ bilinearly resizes the prior feature to the corresponding spatial resolution. Here E_s denotes the feature after the initial embedding or preceding downsampling operation and before the restoration blocks at level s. The affine maps modulate this feature:

$$\widetilde{E}_s = E_s \odot [1 + a \tanh(\gamma_s)] + a \tanh(\beta_s), \quad a = 0.1. \tag{20}$$

The three encoder levels have 24, 48, and 96 channels. Each affine head outputs maps with the same channel dimension as its encoder input. Its final convolution is initialized to zero, making the initial modulation an identity mapping.

B.2 Cross-Modulation Across Feature Depths and Frequencies

Extraction at multiple feature depths. We extract normalized patch-token maps from zero-based blocks $\mathcal{S} = \{0, 2, 4, 6, 8, 10\}$ of the frozen DINOv3 ViT-S/16. The regularly spaced blocks cover early, intermediate, and late representations. The input is clamped to $[0, 1]$, bicubically resized to 224×224, and normalized using the encoder's ImageNet statistics. A trainable bias-free 1×1 projection P_l maps each token feature $Z_l(x)$ from 384 channels to 24 channels.

For a feature map A, we define the band operator

$$\begin{gathered} \mathcal{B}_b(A) = \mathrm{irFFT2}(M_b \odot \mathrm{rFFT2}(A)), \\ b \in \mathcal{B} = \{L, M, H\}. \end{gathered} \tag{21}$$

The masks in Eq. (12) are evaluated on the native frequency grid of A using the normalized radial coordinate in Eq. (11). Both transforms use orthonormal normalization, and the inverse transform restores the original spatial dimensions. The three reconstructed components sum to A up to numerical precision.

Band-specific feature-depth aggregation. Each frequency band learns a separate distribution over the selected feature depths:

$$\begin{aligned} \pi_{b,l} &= \frac{\exp(\theta_{b,l})}{\sum_{k \in \mathcal{S}} \exp(\theta_{b,k})}, \\ D_b(x) &= \mathcal{U}\left(\sum_{l \in \mathcal{S}} \pi_{b,l} \mathcal{B}_b(P_l Z_l(x)) \right), \end{aligned} \tag{22}$$

where $\mathcal{U}$ bilinearly resizes the aggregate to the coarsest decoder resolution. Projection, decomposition, and feature-depth aggregation are all performed at the native 14×14 token resolution before resizing.

The parameters θ are initialized to zero, so that $\pi_{b,l} = 1/6$. These feature-depth weights are optimized jointly with the restoration network and are shared across images. Input-dependent mixing weights are introduced only after this aggregation step, when the router assigns weights to the three frequency branches.

Same-band modulation. Let $F \in \mathbb{R}^{96 \times H/4 \times W/4}$ denote the output of E_3 after its restoration blocks and before visual-prior injection and decoder stage D_3. A band-specific 1×1 adapter C_b maps $D_b(x)$ to the channel dimension of F and produces the modulation

$$G_b = \tanh(C_b(D_b(x))) \odot \mathcal{B}_b(F). \tag{23}$$

The DINOv3 and backbone features use the same relative frequency regions on their respective spatial grids. The modulation varies spatially, whereas the router assigns a single image-level weight to each branch. Because spatial multiplication can introduce cross-frequency interactions, G_b should be interpreted as a band-conditioned contribution rather than a strictly band-limited output.

B.3 Dynamic Frequency Routing

As shown in Figure 2(d), the router summarizes the three guidance maps $\{D_b(x)\}_{b\in\mathcal{B}}$ and predicts a distribution $p(x)$ over the frequency branches. A bounded transformation then converts this distribution into the injection weights $w(x,t)$ used by the restoration network. During training, the predictor is supervised using a frequency-discrepancy target derived from the paired reference.

Frequency-statistic descriptor. For each $D_b(x)$, we compute its RMS magnitude r_b, mean absolute magnitude m_b, and population standard deviation s_b over channels and spatial positions:

$$\begin{aligned} r_b &= \sqrt{\langle D_b^2\rangle + 10^{-12}}, \\ m_b &= \langle |D_b|\rangle + 10^{-12}, \\ s_b &= \mathrm{Std}_{\mathrm{pop}}(D_b) + 10^{-12}, \end{aligned} \tag{24}$$

where $\langle\cdot\rangle$ denotes a per-image mean. Before taking its logarithm, r_b is additionally floored at 10^{-8}. The nine statistics are processed as

$$\begin{aligned} d(x) &= \mathrm{LN}([\log \mathbf{r}, \log \mathbf{m}, \log \mathbf{s}]), \\ z(x) &= \mathrm{clip}(f_\phi(d(x)), -3, 3), \\ p(x) &= \mathrm{softmax}(z(x)/T). \end{aligned} \tag{25}$$

Each vector follows the low/mid/high band order. LN denotes LayerNorm with learned affine parameters, and f_ϕ is a $9 \to 32 \to 3$ MLP with SiLU activations. We set $T = 1.5$. This descriptor captures the magnitude and variation of the three branches using only information from the low-light input.

Reference-derived routing supervision. During training, we derive a routing target from the relative LQ–GT discrepancy in each RGB frequency band. Applying $\mathcal{B}_b$ to the training pair on its native spatial grid gives

$$\begin{aligned} e_b(x,y) &= \frac{\langle |\mathcal{B}_b(y) - \mathcal{B}_b(x)|\rangle}{\langle |\mathcal{B}_b(y)|\rangle + 10^{-3}}, \\ q(x,y) &= \mathrm{sg}[\mathrm{softmax}((e - \bar{e}\mathbf{1})/T_u)], \end{aligned} \tag{26}$$

where $e = (e_L, e_M, e_H)$, $\bar{e} = \frac{1}{3}\sum_b e_b$, $T_u = 0.25$, and sg denotes stop-gradient. The routing objective is

$$\mathcal{L}_{\mathrm{route}} = -\mathbb{E}_{x,y} \sum_{b\in\mathcal{B}} q_b(x,y) \log p_b(x). \tag{27}$$

A larger normalized discrepancy gives a larger target probability. Normalization by the reference magnitude reduces sensitivity to absolute band amplitude, while the offset stabilizes low-energy bands. The resulting q serves as a surrogate for relative repair demand rather than an optimal routing label or a measure of DINOv3 feature reliability. The reconstruction objectives determine how the routed guidance is ultimately used. The reference image contributes only to the construction of the training target; prediction of $p(x)$ depends solely on x.

Bounded frequency allocation. We anchor the predicted distribution to $u = (1/3, 1/3, 1/3)$ and obtain the injection weights as

$$\begin{aligned} c(x) &= 1 - \frac{\mathcal{H}(p(x))}{\log 3}, \\ \mathcal{H}(p) &= -\sum_b p_b \log p_b, \\ \lambda_t &= 0.2\min(t/30000, 1), \\ w(x,t) &= u + \lambda_t c(x)[p(x) - u]. \end{aligned} \tag{28}$$

Here, t is the number of completed optimizer updates before the current forward pass. The first update therefore uses uniform injection, although $\mathcal{L}_{\text{route}}$ already provides supervision to the predictor. The input-dependent correction is introduced gradually over the first 30,000 updates.

The entropy factor $c(x)$ measures the concentration of $p(x)$ and suppresses the correction when the prediction is close to uniform. Since $0 \le \lambda_t c(x) \le 0.2$, the injection weights remain on the probability simplex and satisfy $4/15 \le w_b \le 7/15$. This keeps all three branches active. The distribution $p(x)$ determines the direction of the correction, while $\lambda_t c(x)$ controls its magnitude.

Residual injection. The three branch contributions are combined according to the predicted weights:

$$R_w = \sum_{b\in\mathcal{B}} w_b G_b, \qquad R_u = \frac{1}{3}\sum_{b\in\mathcal{B}} G_b. \tag{29}$$

Changing the mixture can also change the magnitude of the resulting residual. To reduce this effect, we rescale R_w using the uniform mixture R_u constructed from the same branch contributions. With $\rho(R) = \sqrt{\langle R^2\rangle + 10^{-12}}$, $\rho_w = \rho(R_w)$, and $\rho_u = \rho(R_u)$, the injection is

$$\begin{aligned} \alpha &= \mathrm{sg}\left(\frac{\rho_u}{\max(\rho_w, 10^{-12})}\right), \\ \widetilde{R} &= \begin{cases} \alpha R_w, & \rho_w > \max(10^{-4}\rho_u, 10^{-12}), \\ R_u, & \text{otherwise}, \end{cases} \\ F' &= F + \eta\widetilde{R}, \qquad \eta = 0.1\sigma(\xi). \end{aligned} \tag{30}$$

The scalar ξ is initialized to zero, giving $\eta = 0.05$. The fallback handles cases in which the routed residual nearly cancels, and R_u is used directly when $\lambda_t = 0$. The scale α is detached, while gradients through R_w remain active in the routed branch. This operation controls the magnitude of the aggregate residual without normalizing the branches individually.

Let S_1 and S_2 denote the outputs of encoder stages E_1 and E_2, respectively. Starting from F', decoder stage D_3 applies four restoration blocks. Its output is upsampled and fused with S_2 at F_2 before the two restoration blocks of D_2. A second upsampling is fused with S_1 at F_1 before the single restoration block of D_1. Four refinement blocks and the final 3×3 convolution produce the RGB residual Δx:

$$\Delta x = \mathcal{D}(F'; S_1, S_2), \qquad \hat{y} = x + \Delta x, \tag{31}$$

Here $\mathcal{D}$ denotes the decoder, both skip fusions, refinement, and final RGB projection.

B.4 Training Objectives and Inference

Image reconstruction. We optimize pixel-domain and Fourier-domain reconstruction losses:

$$\begin{aligned} \mathcal{L}_{\text{pix}} &= \mathbb{E}_{x,y}\langle|\hat{y} - y|\rangle, \\ \mathcal{L}_{\text{fft}} &= \mathbb{E}_{x,y}\left\langle\left|\mathcal{C}(\mathcal{F}(\hat{y})) - \mathcal{C}(\mathcal{F}(y))\right|\right\rangle. \end{aligned} \tag{32}$$

Here, $\mathcal{F}$ denotes the unnormalized full 2D FFT, and $\mathcal{C}(A) = [\mathrm{Re}\, A, \mathrm{Im}\, A]$ concatenates the real and imaginary components. The Fourier loss averages their elementwise L1 differences. This FFT uses a different normalization from the orthonormal transforms used for feature decomposition.

Cross-exposure guidance alignment. During training, frozen DINOv3 features are also extracted from the reference image y. Let $\bar{D}_b(x)$ and $\bar{D}_b(y)$ denote the aggregates in Eq. (22) before spatial resizing. With $v(\cdot)$ denoting flattening followed by ℓ_2 normalization, we minimize

$$\begin{aligned} \mathcal{L}_{\text{align}} = \mathbb{E}_{x,y}\frac{1}{3}\sum_{b\in\mathcal{B}} \\ \left[1 - v(\bar{D}_b(x))^\top \mathrm{sg}(v(\bar{D}_b(y)))\right]. \end{aligned} \tag{33}$$

The LQ and reference paths share the same projections and feature-depth weights. The reference aggregate is computed using their current parameters and is then detached. This loss encourages cross-exposure consistency in the projected and aggregated guidance at the native 14×14 token resolution.

Joint optimization. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{pix}} + 0.1\mathcal{L}_{\text{fft}} + 0.1\mathcal{L}_{\text{align}} + 0.05\mathcal{L}_{\text{route}}. \tag{34}$$

The restoration backbone and guidance modules are trained jointly from scratch, while the pretrained DINOv3 encoder remains frozen. Through the descriptor, the routing loss updates the predictor, projections, and feature-depth weights. Once $\lambda_t > 0$ and the routed residual is active, reconstruction gradients also propagate through $p(x)$ and $c(x)$. The routing target q, the reference alignment branch, and the residual scale α remain detached.

For Synthetic VisDrone, training runs for 300,000 optimizer updates with AdamW, an initial learning rate of 10^{-3}, weight decay of 10^{-3}, and momentum parameters $(0.9, 0.9)$. The learning rate follows cosine decay to 10^{-7}, and the global gradient norm is clipped to 0.01. We use 128×128 crops and an effective batch size of 32.

Inference. Inference requires only the low-light input x. The photometric branch, frozen DINOv3 encoder, frequency aggregation, router, and residual injection remain active, with the completed warmup value $\lambda = 0.2$. Reference feature extraction and all auxiliary losses are disabled.

C Dataset and Evaluation Protocols

Datasets. We evaluate DPFR-LLIE on Synthetic VisDrone Low-Light and LOL-Blur with paired references, and on DarkRS using unpaired nighttime UAV images. Synthetic VisDrone Low-Light and LOL-Blur follow separate training and evaluation protocols. Transfer to DarkRS is evaluated using models trained on LOL-Blur.

We construct Synthetic VisDrone Low-Light from normally exposed VisDrone images. Motivated by inverse-curve darkening and noise simulation for low-light remote sensing (Yao et al., 2024), the synthesis combines an inverse DCE curve transformation (Li et al., 2021) with Gaussian noise and motion blur; motion blur is a UAV-oriented extension. The degraded image I_{low} and its reference I_{gt} form a pair. The manifest contains 19,246 training pairs and 2,059 validation pairs across light, medium, heavy, and extreme degradation.

Recorded synthesis parameters. Table 6 details the four severity settings recorded in the pair manifest. The exposure targets and accepted intervals are given by target_exp, exp_min, and exp_max. The curve column lists the two endpoints in curve_a_range, reordered from lower to upper value; it is distinct from the per-image scale recorded as a_scale_selected. Observed scale values occur in increments of 0.1 within the listed ranges. The Gaussian-noise standard deviations are recorded in 8-bit units, equivalent to $6/255$, $10/255$, $16/255$, and $22/255$ on a normalized intensity scale. Motion lengths list the discrete recorded values; their frequencies are not assumed to be uniform. Recorded motion angles span approximately $0.061°$–$179.986°$. The final-brightness ranges summarize final_mean_y; all 21,305 entries lie within their corresponding exposure intervals and carry an in_range exposure status. The table distinguishes fixed recorded settings from realized per-image parameter and brightness ranges.

Source images and split accounting. The 21,305 pairs correspond to 5,446 distinct GT filenames, with 4,918 in training and 528 in validation. Multiple severity variants can share a GT image within a split, so the pair count is not a count of distinct source images. No GT filename occurs in both splits. This filename-level check does not establish disjoint source videos or scenes. All pair counts and parameter summaries above are computed from the supplied manifest; they do not assume that every source image has all four severity variants.

For LOL-Blur (Zhou et al., 2022), models are trained on the official training split and evaluated on the test split. The resulting models are then applied directly to DarkRS (Yao et al., 2024) without

Table 6: Recorded synthesis parameters and pair counts from pairs_clean_weighted.csv. Curve endpoints are ordered numerically; scale, blur-length, and final-brightness ranges summarize values present in the manifest.

Parameter	Light	Medium	Heavy	Extreme
Training pairs	4,549	4,911	4,914	4,872
Validation pairs	502	520	523	514
Target exposure	0.32	0.22	0.14	0.08
Exposure interval	$[0.28, 0.36]$	$[0.18, 0.26]$	$[0.10, 0.17]$	$[0.06, 0.11]$
Curve endpoints	$[-0.20, -0.10]$	$[-0.34, -0.20]$	$[-0.46, -0.30]$	$[-0.58, -0.42]$
Selected scale range	$[0.2, 1.3]$	$[0.2, 1.3]$	$[0.3, 1.3]$	$[0.4, 1.3]$
Noise std. (8-bit)	6	10	16	22
Motion lengths	$\{3, 5\}$	$\{5, 7\}$	$\{7, 9, 11\}$	$\{9, 11, 13, 15\}$
Final mean brightness	0.2803–0.3598	0.2013–0.2576	0.1232–0.1697	0.0771–0.1100

fine-tuning. DarkRS contains real nighttime UAV images without paired references and is used to evaluate transfer beyond the LOL-Blur training distribution.

Evaluation Metrics. We report PSNR and SSIM on the paired Synthetic VisDrone Low-Light and LOL-Blur datasets, where higher values indicate better reconstruction quality. On DarkRS, we report NIQE, for which lower values are preferred. The no-reference evaluation on DarkRS is complemented by visual comparisons. All ablation variants use the same fixed evaluation pipeline and image list.

Implementation Details. DPFR-LLIE is implemented in PyTorch using BasicSR. All compared methods use the same dataset splits and data-augmentation strategy. For DPFR-LLIE, training images are randomly cropped into 128×128 patches with an effective batch size of 32. The restoration network is trained from scratch, while the DINOv3 ViT-S/16 encoder remains frozen throughout training. We use AdamW with an initial learning rate of 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.9$, and weight decay 10^{-3}. The learning rate follows a cosine schedule over 300k optimizer updates and decreases to 10^{-7}. Separate models are trained for Synthetic VisDrone Low-Light and LOL-Blur, and the LOL-Blur-trained models are used directly for DarkRS without additional training.

Ablation Protocol. All ablation runs use the same Synthetic VisDrone split, restoration-network initialization, frozen DINO checkpoint, crop size, effective batch size, optimizer, learning-rate schedule, and 300k-update training budget. The DINO-only experiment uses four GPUs with a per-GPU batch size of 8. Each guidance-construction control uses a per-GPU micro-batch of 4 with gradient accumulation over 8 steps, giving an effective batch size of 32. Intermediate checkpoints are not used for model selection. Every variant is evaluated at the fixed final 300k checkpoint using the designated evaluation scripts. This fixed-endpoint protocol ensures that differences in validation frequency do not affect which checkpoint is reported.

C.1 Downstream Object Detection Protocol and Error Analysis

Operating-point diagnostic protocol. For the diagnostic visualization and count analysis only, predictions with confidence scores of at least 0.5 are matched to same-class ground-truth boxes at an IoU threshold of 0.5. Matched predictions are marked as true positives (TP), unmatched predictions as false positives (FP), and unmatched ground-truth instances as false negatives (FN). This single operating-point diagnostic is distinct from COCO AP, which evaluates the ranked predictions and averages performance over IoU thresholds from 0.50 to 0.95.

Why visually favorable examples can still yield modest AP. Figure 6 shows that DPFR-LLIE converts more missed objects into correct detections than the compared inputs, particularly in dense and severely degraded regions. The full-set diagnostics in Table 7 support this observation: com-

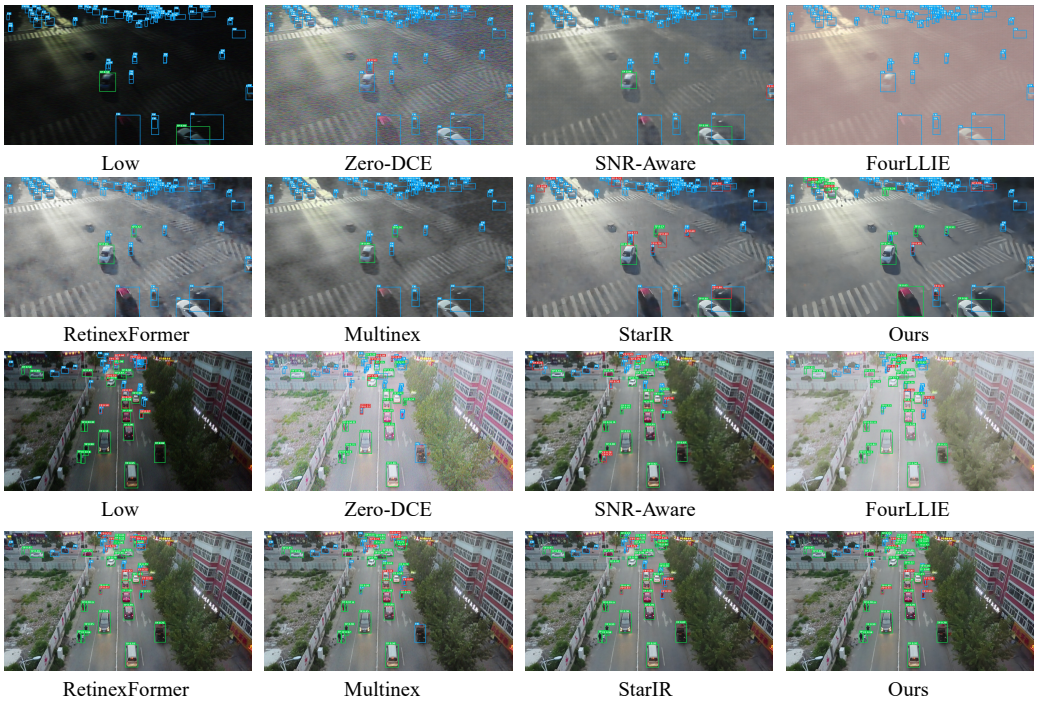

Figure 6: Error-coded detection comparisons on two selected Synthetic VisDrone Low-Light scenes. All panels use the same frozen Faster R-CNN detector; only the detector input changes. Green boxes denote true positives, red boxes denote false positives, and cyan boxes denote false negatives under the diagnostic setting of confidence ≥ 0.5 and IoU ≥ 0.5. Numbers attached to predicted TP and FP boxes are confidence scores. DPFR-LLIE recovers more valid objects than the compared enhancement methods in these examples, while the remaining cyan boxes expose persistent missed detections. These selected cases are diagnostic examples rather than a substitute for the complete 2,059-image COCO evaluation in Table 5.

Table 7: Fixed-operating-point detection diagnostics over all 2,059 validation images. Predictions use confidence ≥ 0.5 and same-class matching at IoU ≥ 0.5. These values diagnose TP/FP/FN behavior and are not COCO AP. Bold indicates the best value in each column.

Input	TP	FP	FN	P (%)	R (%)	F_1 (%)
Low input	33,380	11,229	113,577	74.83	22.71	34.85
Zero-DCE	25,669	9,882	121,288	72.20	17.47	28.13
SNR-Aware	36,480	16,720	110,477	68.57	24.82	36.45
FourLLIE	25,609	**8,099**	121,348	**75.97**	17.43	28.35
RetinexFormer	45,936	16,948	101,021	73.05	31.26	43.78
Multinex	43,049	15,096	103,908	74.04	29.29	41.98
StarIR	57,586	21,276	89,371	73.02	39.19	51.00
DPFR-LLIE	**58,527**	21,395	**88,430**	73.23	**39.83**	**51.59**

pared with StarIR, DPFR LLIE produces 941 additional true positives and 941 fewer false negatives, at the cost of 119 additional false positives.

Nevertheless, 88,430 objects remain undetected and the diagnostic recall is only 39.83%. Most remaining errors occur on small, densely distributed, or severely degraded targets, consistent with the relatively low absolute AP_S of 8.41. Moreover, a box counted as a TP at IoU 0.5 in the visualization may fail at stricter localization thresholds. Because COCO AP averages over IoU thresholds

from 0.50 to 0.95 and considers the complete confidence ranking, clearer appearance and favorable selected examples do not necessarily translate into a high absolute AP. The qualitative results should therefore be interpreted together with, rather than in place of, the complete-set metrics.

D Detailed Ablation Controls

We use Synthetic VisDrone Low-Light to examine three aspects of the design: the complementary roles of the photometric and DINO priors, the effect of input-dependent frequency allocation, and the construction of frequency-specific guidance across feature depths. Except for the component under study, all variants use the same restoration backbone, initialization, frozen DINO encoder, training budget, and fixed-final-checkpoint evaluation protocol.

Contributions of the Two Priors. Table 2 compares the backbone with the Multinex-style photometric prior, DINO guidance, and their combination.

Adding the photometric prior improves the backbone from 26.10/0.853 to 26.69/0.865 in PSNR/SSIM. DINO guidance alone also improves on the backbone, indicating that frozen structural features provide useful information without the photometric branch. Combining the two priors gives the best overall result, showing that image-derived photometric modulation and DINO-derived structural guidance provide complementary information. Because the DINO branch includes its alignment and routing objectives, this ablation measures the complete branch together with its associated supervision rather than the frozen encoder alone.

Effect of Dynamic Routing. Table 3 compares uniform, learned static, and input-dependent injection policies.

Uniform routing assigns equal weights to the three frequency branches. Static routing learns a single mixture shared across images, whereas dynamic routing predicts a separate mixture for each input. The static and dynamic variants use the same uniform anchor, confidence transformation, routing-strength schedule, and residual-scale control. For the static variant, global logits define p_s and c_s, giving $w_s = u + \lambda_t c_s(p_s - u)$; nonuniform initialization is used to avoid the stationary uniform point. In the uniform and static variants, the image-conditioned predictor and routing loss remain active as an auxiliary branch. They continue to supervise the shared guidance features but do not determine the injected mixture.

Dynamic routing achieves 26.81 dB PSNR, compared with 26.73 dB for uniform routing and 26.75 dB for static routing. The corresponding gains are 0.08 and 0.06 dB. Its SSIM of 0.867 matches uniform routing and is 0.001 below the static variant. These matched injection-policy comparisons support a modest empirical PSNR advantage for input-conditioned routing under the evaluated protocol, while the SSIM differences remain limited. Equal-RMS matching keeps the aggregate residual magnitude comparable across routing policies, reducing the influence of residual scale on this comparison.

Frequency Decomposition and Feature-Depth Aggregation. Table 4 in the main paper reports the frequency and feature-depth aggregation controls. We provide their detailed definitions below.

The broadband variant replaces frequency filters with identity operators in the guidance branches and auxiliary objectives, retaining three modulation branches and separate feature-depth mixtures. This changes both the representation and its supervision. In particular, the three RGB discrepancies in Eq. (8) become identical, making the routing target uniform. The 0.11 dB PSNR difference therefore evaluates the combined frequency-structured representation and supervision design.

Single feature depth uses zero-based block 10. Uniform fusion assigns weights $1/6$ to the selected six blocks; shared fusion learns one mixture for all bands. Full achieves higher PSNR than these tested configurations, including 0.06 dB over shared fusion. The single-feature-depth result refers specifically to the block-10 configuration. Shared-versus-uniform fusion also involves parameterization and optimization: because P_l is trainable and linear, a shared mixture π_l can be absorbed into $P'_l = 6\pi_l P_l$ for uniform aggregation. This equivalence of the aggregation mapping does not imply identical optimization or regularization. Band-specific weights generally cannot be absorbed into one common projection in the same way, distinguishing the full formulation from shared fusion.

Ablation Summary. The guidance-construction comparisons favor the full band-specific configuration among the evaluated alternatives, and Dynamic yields 0.08/0.06 dB higher PSNR than Uniform/Static. At the system level, the photometric prior raises PSNR from 26.10 to 26.69 dB; adding the complete DINO branch and its associated supervision raises it to 26.81 dB. DINO guidance alone achieves 26.47 dB. These conditional gains summarize the contributions of guidance construction, allocation, and prior integration within their respective comparison settings.

E Additional Quantitative and Qualitative Results

We compare DPFR-LLIE with Zero-DCE (Li et al., 2021), SNR-Aware (Xu et al., 2022), FourLLIE (Wang et al., 2023), RetinexFormer (Cai et al., 2023), Multinex (Brateanu et al., 2026), and StarIR (Cui et al., 2026). These methods cover zero-reference enhancement, SNR-aware restoration, frequency-domain processing, Retinex-based enhancement, and the architecture that informs our restoration path.

Quantitative Comparison on Synthetic VisDrone. Table 8 reports the overall results and those for each degradation level.

Table 8: Quantitative comparison on Synthetic VisDrone Low-Light, reported overall and for each degradation level. Higher PSNR and SSIM are better. Bold indicates the best result in each column.

	Overall		Light		Medium	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Zero-DCE	14.58	0.515	11.36	0.679	14.23	0.626
SNR-Aware	20.86	0.732	22.19	0.831	21.42	0.777
FourLLIE	14.58	0.433	21.71	0.842	13.57	0.571
RetinexFormer	24.15	0.809	27.83	0.922	25.20	0.866
Multinex	22.06	0.753	24.95	0.906	22.98	0.835
StarIR	26.10	0.853	30.77	0.952	27.40	0.907
DPFR-LLIE	**26.81**	**0.867**	**31.70**	**0.958**	**28.26**	**0.917**

	Heavy		Extreme	
Method	PSNR	SSIM	PSNR	SSIM
Zero-DCE	16.31	0.461	16.33	0.299
SNR-Aware	20.47	0.704	19.39	0.617
FourLLIE	11.75	0.221	11.51	0.112
RetinexFormer	22.77	0.778	20.92	0.674
Multinex	21.11	0.708	19.26	0.566
StarIR	24.27	0.826	22.08	0.727
DPFR-LLIE	**24.85**	**0.841**	**22.56**	**0.744**

DPFR-LLIE achieves 26.81 dB PSNR and 0.867 SSIM, improving on StarIR by 0.71 dB and 0.014 SSIM. It obtains the highest PSNR and SSIM among the compared methods at all four degradation levels. Relative to StarIR, the PSNR gains from light to extreme degradation are 0.93, 0.86, 0.58, and 0.48 dB, respectively. Under heavy and extreme degradation, DPFR-LLIE achieves 24.85/0.841 and 22.56/0.744 in PSNR/SSIM. The improvement over the backbone therefore remains consistent across the evaluated degradation levels.

Qualitative Comparison on Synthetic VisDrone. Figure 3 shows two UAV scenes under extreme degradation.

The compared methods exhibit different trade-offs among illumination correction, noise suppression, and detail preservation. In the examples shown, DPFR-LLIE improves the visibility of vehi-

cles, road markings, and local boundaries while reducing the color cast in the degraded inputs. The enlarged regions provide a closer comparison of local reconstruction against the paired references.

LOL-Blur Evaluation and Transfer to DarkRS. Table 9 reports the results of models trained on LOL-Blur.

Table 9: LOL-Blur test results and transfer to DarkRS. All models in this table are trained on LOL-Blur and evaluated on DarkRS without fine-tuning. Bold indicates the best result in each column.

	LOL-Blur		DarkRS
Method	PSNR↑	SSIM↑	NIQE↓
Zero-DCE	14.95	0.612	5.072
SNR-Aware	21.87	0.811	9.205
FourLLIE	18.53	0.732	4.130
RetinexFormer	24.00	0.840	5.440
Multinex	22.82	0.823	**3.959**
StarIR	26.52	0.905	4.184
DPFR-LLIE	**27.94**	**0.927**	4.129

On the LOL-Blur test split, DPFR-LLIE achieves 27.94 dB PSNR and 0.927 SSIM, improving on StarIR by 1.42 dB and 0.022 SSIM. When applied directly to DarkRS, DPFR-LLIE obtains a NIQE score of 4.129, compared with 3.959 for Multinex, 4.130 for FourLLIE, and 4.184 for StarIR. These results show strong paired reconstruction performance on LOL-Blur while maintaining competitive no-reference quality when transferred to real nighttime UAV images.

Visual Comparison on LOL-Blur and DarkRS. Figure 7 presents examples from LOL-Blur and DarkRS.

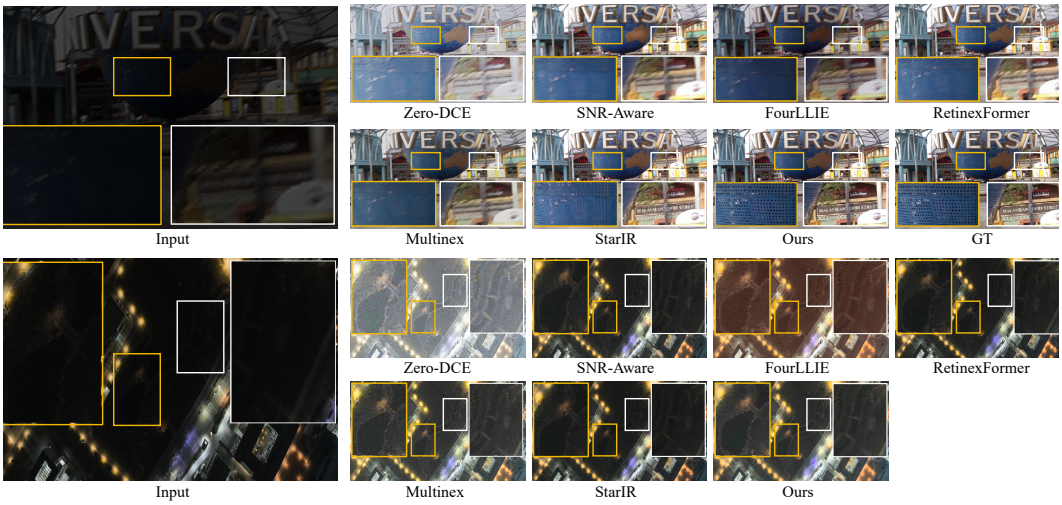

Figure 7: Qualitative comparison on LOL-Blur and DarkRS using models trained on LOL-Blur. The LOL-Blur example includes a paired reference, whereas the DarkRS example is a real nighttime UAV image without a paired reference. Enlarged regions highlight local structures.

In the LOL-Blur example, the paired reference allows direct comparison of reconstructed lettering and building structures. In the DarkRS example, DPFR-LLIE improves visibility in dark regions around illuminated roads and buildings. These examples complement the paired reconstruction metrics on LOL-Blur and the no-reference evaluation on DarkRS.